IRE Transactions

ON AUTOMATIC CONTROL

Volume AC-4 **NOVEMBER, 1959** **Number 2**

TABLE OF CONTENTS

The Following Papers Are Reprinted From
Part 4 of the 1959 IRE NATIONAL CONVENTION RECORD

The Following Papers Are Reprinted From
Part 4 of the 1959 IRE WESCON CONVENTION RECORD

PUBLISHED BY THE

PROFESSIONAL GROUP ON AUTOMATIC CONTROL

IRE TRANSACTIONS®
on Automatic Control

George S. Axelby, *Editor*, Air Arm Division, Westinghouse Electric Corp., Box 746, Baltimore, Md.

Published by the Institute of Radio Engineers, Inc., for the Professional Group on Automatic Control, 1 East 79th Street, New York 21, New York. Responsibility for the contents rests upon the authors, and not upon the IRE, the Group or its members. Individual copies available for sale to IRE-PGAC members at $0.85, to IRE members at $1.25, and to nonmembers at $2.55.

ON ADAPTIVE CONTROL PROCESSES

Richard Bellman and Robert Kalaba
The RAND Corporation
Santa Monica, California

Summary

One of the most challenging areas in the field of automatic control is the design of automatic control devices that 'learn' to improve their performance based upon experience, i.e., that can adapt themselves to circumstances as they find them. The military and commercial implications of such devices are impressive, and interest in the two main areas of research in the field of control, the USA and the USSR, runs high. Unfortunately, though, both theory and construction of adaptive controllers are in their infancy, and some time may pass before they are commonplace. Nonetheless, development at this time of adequate theories of processes of this nature is essential.

The purpose of our paper is to show how the functional equation technique of a new mathematical discipline, dynamic programming, can be used in the formulation and solution of a variety of optimization problems concerning the design of adaptive devices. Although, occasionally, a solution in closed form can be obtained, in general, numerical solution via the use of high-speed digital computers is contemplated.

We discuss here the closely allied problems of formulating adaptive control processes in precise mathematical terms and of presenting feasible computational algorithms for determining numerical solutions.

To illustrate the general concepts, consider a system which is governed by the inhomogeneous Van der Pol equation

$$\ddot{x} + \mu(x^2 - 1)\dot{x} + x = r(t), \quad 0 \leq t \leq T,$$

where $r(t)$ is a random function whose statistical properties are only partially known to a feedback control device which seeks to keep the system near the unstable equilibrium state $x = 0$, $\dot{x} = 0$. It proposes to do this by selecting the value of μ as a function of the state of the system at time t, and the time t itself. By observing the random process $r(t)$, the controller may, with the passage of time, infer more and more concerning the statistical properties of the function $r(t)$ and thus may be expected to improve the quality of its control decisions. In this way the controller adapts itself to circumstances as it finds them. The process is thus an interesting example of adaptive control, and, conceivably, with some immediate applications.

Lastly, some areas of this general domain requiring additional research are indicated.

I. Introduction

In many engineering, economical, biological, and statistical control processes, a device of one type or another which we shall call a controller is called upon to perform under various conditions of uncertainty regarding the structure of the underlying physical processes. These conditions range all the way from complete knowledge to total ignorance. As the process unfolds, however, additional information concerning these factors may become available to the controller, which then has the possibility of 'learning' to improve its performance based upon experience, or in fact, actual experimentation. In this case we say that the controller adapts itself to its environment.

In an earlier paper, [7], a broad and general foundation was laid for the mathematical study of adaptive processes, through the use of concepts from the field of dynamic programming, [4]. The specific purpose of this paper is to render these notions more concrete through the detailed study of some special control processes involving a nonlinear system with a tendency to be stimulated to undesirable oscillations.

We approach the adaptive control process in a series of steps. First we assume that the controller has complete information concerning the behavior of the forcing function over time, a process which is referred to as a deterministic control process. Then we introduce some unknown factors, which appear mathematically as random variables having distribution functions which are known to the controller. This leads to a stochastic control process. Lastly, we allow the controller still less information about the unknown factors and require that the controller learn to improve its performance through observation of the values of $r(t)$, an adaptive control process.

In this paper, we limit the deficiency in the controller's knowledge to incomplete information concerning a random disturbing force. There are, needless to say, many other ways in which ignorance can manifest itself. Among these we may mention uncertainties concerning the determination of the state of the system and its environment by the sensing devices, the objective (figure of merit) of the process, the transformations of the state of the system produced by control decisions, the set of allowable decisions, and so on. These will be examined in subsequent investigations.

Although the design and operation of adaptive controllers are in their infancy, interest in these devices runs high, [10]. The functional equation technique of dynamic programming, (4), can be used to attack a wide variety of problems involving the determination of optimal control policies for control devices having the ability to adapt themselves to circumstances. In particular, it provides a useful conceptual framework for the very discussion of such devices. Though on occasion, analytic results are obtained, [?), emphasis is upon the development of methods which are suitable for use in conjunction with high-speed digital computers having large memories.

Some of the advantages of the dynamic programming approach are its suitability for use with nonlinear, as well as linear, systems, its automatic production of a desirable parameter study (a 'sensitivity' or stability analysis), its straightforwardness and computational feasibility, and its ability to incorporate stochastic elements in a routine fashion.

Let us now turn our attention to an example which will serve to illustrate these remarks.

2. A Feedback Control Problem

Let us consider a system which, if uncontrolled, is governed by the well-known nonlinear differential equation

$$\ddot{x} + \mu(x^2 - 1)\dot{x} + x = 0, \; x(0) = c_1, \; \dot{x}(0) = c_2, \quad (1)$$

(the Van der Pol equation) where the dots denote differentiation with respect to time. This equation is of fundamental importance in describing the development of relaxation oscillations in triode oscillator circuits and in describing the operation of multivibrators. We shall call μ the system parameter. If we introduce the function $v(t)$, by means of the relation

$$\dot{x} = v, \quad (2)$$

then the equation in (1) can be replaced by the first-order system

$$\dot{x} = v, \; x(0) = c_1, \quad (3)$$

$$\dot{v} = -\mu(x^2 - 1)v - x, \; v(0) = c_2.$$

It is well-known that if $c_1^2 + c_2^2 \neq 0$, then the solution of the system (3) will tend toward a unique periodic solution. In the (x,v) phase plane, this periodic solution is represented by a closed curve which all trajectories (except $x = 0$, $v = 0$) approach. Thus, when the system is disturbed from its (unstable) equilibrium position ($x = 0$, $v = 0$), a periodic oscillation tends to develop. Full details are available in the book on nonlinear oscillations by Stoker, [1].

Let us assume, though, that the oscillations are undesirable ('parasitic'), and that the system can be controlled by varying the system parameter, μ, in a given range in an effort to maintain the system as close as possible to its equilibrium state.

Consider that the process begins at time 0 and terminates at time T, and that the system is initially in state (c_1, c_2), where c_1 is the displacement, x, and c_2 is the velocity, $\dot{x} = v$.

We shall arbitrarily measure the 'cost' of deviation from equilibrium during the process by the integral

$$J[\mu] = \int_0^T (|x(t)| + |v(t)|)dt + \exp(|x(T)| + |v(t)|), \quad (4)$$

where $\exp(z)$ is the exponential function of z. We deliberately use such a monstrous function in order to squelch any direct analytic approach in embryo. Our objective will be the determination of the system parameter μ as a function of the state of the system at time t and the time t itself, for $0 \leq t \leq T$ in order to minimize $J[\mu]$. The control function μ will be subject to a constraint $m_1 \leq \mu(t) \leq m_2$, where m_1 may be negative. Notice that the criterion function is not the usual mean-square deviation, which in this case would be of little avail since the underlying equations are nonlinear. The first term on the right-hand side of Eq. (4) measures the cost of deviation during the entire course of the process and the second term measures the cost of deviation at the termination of control.

The temptation is to view this as a problem in the calculus of variations, [16], in which one seeks to determine $\mu = \phi(t)$ as a function of time over the entire interval $0 \leq t \leq T$ in an attempt to minimize the functional J. The fact that μ is constrained to lie between m_1 and m_2 is a cause of some complication. Furthermore, there are no classical prototypes for the stochastic control processes we wish to study below.

The approach which we shall use, by contrast, emphasizes the feedback control nature of the problem. We shall imbed the original problem within a class of problems in which we regard the system as being in some general state $x = c_1$, $\dot{x} = c_2$ at the time t, and ask what the optimal choice of μ is under these circumstances. Notice (as a consequence of the usual existence and uniqueness results for differential equations) that the past history of the process is of no consequence in making this decision, only the current state. Pursuing this approach, in which we have a continuous decision process, [2], we characterize the curve $\mu = \phi(t)$ as an envelope of tangents, rather than as a locus of points, as would be the case were the earlier viewpoint adopted.

In order to prepare the way, though, for the use of digital computing machines, we wish to reformulate the problem in terms of a discontinuous time variable, which will also materially simplify matters conceptually when we deal with the cases of stochastic and adaptive control.

3. A Discrete Version

The problem could be treated in the form in which it now stands, [2]. Since our objective is to devise methods which are particularly suitable for high-speed digital computational purposes, we prefer to reformulate the model itself in terms of discrete variables. It must be borne in mind that, in any event, digital computers consider all variables to be discrete.

The time interval from 0 to T is divided into N intervals of length h so that

$$Nh = T \quad (1)$$

If the system is in state (x_k, v_k) at time kh, and the control decision at that time is to have the system parameter be μ_k, then the new state at time $(k + 1)h$ is given by the finite difference equations

$$\begin{cases} x_{k+1} = x_k + v_k h \\ v_{k+1} = v_k - \left[\mu_k(x_k^2 - 1)v_k + x_k\right]h, \end{cases} \quad (2)$$

relations which hold for $k = 0,1,2,\ldots, N-1$. These equations are the finite difference analogues of the equations in (2.3). The cost of the deviation from equilibrium from time kh to time $(k+1)h$ is taken to be $(|x_k| + |v_k|)h$, and the cost for deviation of the final state from equilibrium is $\exp(|x_N| + |v_N|)$. The total cost for deviation from the equilibrium state during the entire process is considered to be

$$\sum_{k=0}^{N-1} \left[(|x_k| + |v_k|)h \right] + \exp(|x_N| + |v_N|) \quad (3)$$

$$= J\{\mu_0, \mu_1, \ldots, \mu_{N-1}\},$$

the analogue of equation (2.4). Here μ_k is the value of the system parameter selected at time kh. Let us assume that the system is in state (x,y) initially and that we seek a set of parameter values, $[\mu_0, \mu_1, \ldots, \mu_{n-1}]$, which will minimize the total cost of deviation given in equation (3).

4. Deterministic Control

As stated, the problem requires a constrained N-dimensional minimization to be performed, and as such may be quite difficult to carry out computationally in its native form. Even so, this problem is conceptually much simpler than the original continuous version which required a minimization over elements in a function space. To solve this new discrete problem we imbed the given decision process within a class of processes in such a way that we shall have a sequence of N simple one-dimensional optimizations to perform, rather than the one difficult N-dimensional problem. This decomposition makes possible an efficient machine solution.

The imbedding is accomplished by focusing our attention upon determining what value between m_1 and m_2 of the system parameter to choose at time $(N-k)h$ if the system is then in some general state (a,b), where k may have any of the values $0,1,2,\ldots,N-1$ and a and b are any real numbers. The original discrete problem is one of the members of this class of problems. Notice that this is the general problem of interest to the feedback controller, for it must decide what value of the system parameter to call for with the system in some physical state (a,b) and kh time units remaining before the termination of the process.

To formulate the problem analytically, we note first that the minimal cost of deviation over the last k stages of the process with the system starting this portion of the process in some state, say (c_1,c_2), is some definite function of k and c_1 and c_2. It is, namely, the cost that is incurred during the last k stages of the process using an optimal selection of the sequence of system parameter values with the system in state (c_1,c_2) at time $(N-k)h$. Let us therefore define for $k = 1,2,\ldots,N$, the functions

$f_k(c_1,c_2)$ = the cost of the last k stages of the control process with the system beginning those last k stages in state (c_1,c_2), and using an optimal selection of the system parameter throughout the remainder of the process. (1)

We shall determine first $f_1(c_1,c_2)$, then $f_2(c_1,c_2)$, and so on, until $f_N(c_1,c_2)$ has been determined. At the same time, we shall determine the optimal choices of μ to make.

The function $f_1(c_1,c_2)$ is easily determined. Here we are concerned with a process which begins at time $(N-1)h$ and terminates at time $Nh = T$, with the system in state (c_1,c_2) at time $(N-1)h$. The cost of deviation during the process is $(|c_1| + |c_2|)h$. If the value of the state parameter selected is μ, then the state at the termination of the process will be given by the equations

$$x_N = c_1 + c_2 h, \quad (2)$$

$$v_N = c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1 \right] h,$$

where use has been made of the formulas in (3.2). The cost of this terminal deviation is

$$\exp\left[|c_1 + c_2 h| + \left| c_2 - h\left[\mu(c_1^2 - 1)c_2 + c_1 \right] \right| \right].$$

Consequently, the system parameter μ must be selected so that the total cost, given by the expression

$$|c_1| h + |c_2| h + \exp\left[|c_1 + c_2 h| + \left| c_2 - h\left[\mu(c_1^2 - 1)c_2 + c_1 \right] \right| \right],$$

is minimized. The minimizing value of μ for this one-stage process will depend on the state (c_1,c_2), and it can easily be determined by a digital computer using a search technique. The bracket is evaluated for sample values of μ in the range $m_1 \le \mu \le m_2$ and the value of μ yielding the smallest value of the bracket is the optimal system parameter value. Here we see by inspection that μ should be chosen equal to m_2 if $(c_1^2 - 1)c_2 > 0$ and equal to m_1 if $(c_1^2 - 1)c_2 < 0$. If $(c_1^2 - 1)c_2 = 0$, the choice of μ is immaterial. Let us denote this optimal choice of μ for the one-stage process under consideration with the system initially in state (c_1,c_2) by

$$\mu = M_1(c_1,c_2). \tag{3}$$

We have the expression

$$f_1(c_1,c_2) = \min_{m_1 \le \mu \le m_2} \left[|c_1|h + |c_2|h + \exp\left[|c_1 + c_2 h| + |c_2 - h\left[\mu(c_1^2 - 1)c_2 + c_1\right]| \right] \right], \tag{4}$$

where the right-hand side represents the minimum over all choices of μ between m_1 and m_2 of the expression in brackets. It is clear that this minimum value actually depends on c_1 and c_2.

Now let us assume that the functions $f_1(c_1,c_2)$, $f_2(c_1,c_2),\ldots,$ $f_k(c_1,c_2)$, with $k < N$, have all been determined. We wish next to determine the function $f_{k+1}(c_1,c_2)$. We do this by making use of the principle of optimality, [4], which is a special, but quite important, application of the concept of invariant imbedding, [8]. According to this principle, an optimal sequence of decisions has the property that whatever the initial decision and initial state are, the remaining decisions must constitute an optimal sequence of decisions with respect to the state which results from the first decision.

To apply this principle, let us suppose that at time $(N - k - 1)h$, with $k + 1$ decisions remaining and the system in some state (c_1,c_2), the first decision is to set the system parameter equal to μ. The effect of this decision is to transform the system into the state (x_k,v_k) at time $(N - k)h$ (when k decisions remain), where

$$x_k = c_1 + c_2 h,$$

$$v_k = c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right] h. \tag{5}$$

From the cost point of view we see that this results in a cost $(|c_1| + |c_2|)h$ during the time interval from $(N - k - 1)h$ to $(N - k)h$ and (since optimal decisions must be made over the remaining k decisions beginning with the system in the state (x_k,v_k) given by equation (8)), a cost of $f_k(c_1 + c_2 h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right] h)$ during the remainder of the process. Clearly, the choice of the system parameter at time $(N - k - 1)h$ must be made so as to minimize the sum of these two costs. This observation results in the equation

$$f_{k+1}(c_1,c_2) = \min_{m_1 \le \mu \le m_2} \left\{ |c_1|h + |c_2|h + f_k(c_1 + c_2 h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right] h) \right\}, \tag{6}$$

which holds for $k = 1,2,\ldots,N - 1$.

In particular, since $f_1(c_1,c_2)$ is known from the above discussion, $f_2(c_1,c_2)$ can be determined from the formula

$$f_2(c_1,c_2) = |c_1|h + |c_2|h + \min_{m_1 \le \mu \le m_2} \left\{ f_1(c_1 + c_2 h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right] h) \right\}, \tag{7}$$

which follows from equation (6). The value of which minimizes the expression in brackets in equation (7) is the optimal value of μ to choose at time $(N - 2)h$ (the initial stage of a two-stage decision process) with the system in state (c_1,c_2). We denote this optimal choice of μ by

$$\mu = M_2(c_1,c_2). \tag{8}$$

Similarly, the remaining minimal cost functions $f_k(c_1,c_2)$ and optimal decision functions $M_k(c_1,c_2)$ can now be determined recursively.

In order to apply this solution, it is necessary, of course, to construct a control device that will call for the indicated optimal value of the system parameter μ for each state of the system, and time remaining. Should this prove not to be feasible, other sub-optimal policies will have to be employed. These can also be determined by dynamic programming methods, by imposing suitable constraints on the allowable choice of μ and the information fed into the computer-controller. In the event of this sub-optimization, [13], the loss in system performance can be quantitatively assessed by comparison with the performance of an optimal controller.

5. Infinite Duration

In the event that the process is of sufficiently great duration, we may wish to approximate to it by means of an infinitely long process. Furthermore, we may now wish to exert control so as to minimize the maximum deviation of the function $|x(t)| + |v(t)|$ over all time. Let us then define the function

$f(c_1,c_2)$ = the maximum value of $|x|+|v|$ over all time with the system initially in the state (c_1,c_2), using an optimal control policy. (1)

The relevant functional equation now becomes

$$f(c_1,c_2) = \min_{m_1 \le \mu \le m_2} \left\{ \max\left[|c_1| + |c_2|, f(c_1 + c_2 h, c_2 - (\mu(c_1^2 - 1)c_2 + c_1)h) \right] \right\} \quad (2)$$

This is based on the observation that μ must be chosen so that the larger of $|c_1| + |c_2|$ and the greatest deviation over the remainder of the process will be as small as possible. A further discussion of such equations can be found in [2].

6. Stochastic Control

Let us now complicate matters for the controller, somewhat, by assuming that the system is subjected to a random external force, the influence of which cannot be neglected. If we denote the random force which is exerted at time kh by r_k, then equations (3.2) become

$$\begin{aligned} x_{n+1} &= x_n + v_n h, \\ v_{n+1} &= v_n - \left[\mu_n (x_n^2 - 1)v_n + x_n \right] h + r_n h. \end{aligned} \quad (1)$$

The controller can no longer predict precisely what state the system will be transformed into when a value for μ_n is chosen, for the transformed state depends on the value that the random force r_n assumes.

For simplicity, we shall assume that the random variables r_n are independent and that

$$\begin{aligned} \text{Prob}\left\{ r_n = \delta \right\} &= p, \\ \text{Prob}\left\{ r_n = -\delta \right\} &= 1 - p, \end{aligned} \quad (2)$$

so that the disturbing force is either $\pm\delta$. The case where r_n is correlated to the value of r_{n-1} can also be considered at some increase in complexity.

We do, however, wish to assume that the value of p have the known value p^*. This last assumption is not justified in many situations. If the value of p is not known, further complications arise, leading to the adaptive control processes discussed in §7.

Once again we wish to control the development of the system in such a way that

$$J = \sum_{k=0}^{N-1} (|x_k| + |v_k|)h + \exp(|x_N| + |v_N|) \quad (3)$$

is minimized. This now can only be accomplished in some average sense, since x_k and v_k are random variables for $k = 1,2,\ldots,N$. Once again we imbed the original process within a class of processes. Denoting the taking of an expected value by E, we define the sequence of functions

$$f_k(c_1,c_2) = \min_{m_1 \le \mu_i \le m_2} E\left\{ \sum_{i=N-k}^{N-1} (|x_i| + |v_i|)h + \exp(|x_N| + |v_N|) \right\}, \quad (4)$$

for $k = 1,2,\ldots,N$, where

$$x_{N-k} = c_1, \quad v_{N-k} = c_2. \quad (5)$$

Thus, $f_k(c_1,c_2)$ represents the minimal expected total cost of deviation from equilibrium for a process beginning at time $(N-k)h$ and terminating at time $T = Nh$, with the system initially in the state (c_1,c_2).

We first consider the one-stage process which begins at time $(N-1)h$, and terminates at $Nh = T$. We have

$$f_1(c_1,c_2) = \min_{m_1 \le \mu \le m_2} E\left\{ |c_1|h + |c_2|h + \exp\left[|c_1 + c_2 h| + |c_2 - (\mu(c_1^2 - 1)c_2 + c_1)h + r_{N-1}h| \right] \right\}, \quad (6)$$

or, taking the expected value over r_{N-1},

$$\begin{aligned} f_1(c_1,c_2) = {} & |c_1|h + |c_2|h \\ & + \min_{m_1 \le \mu \le m_2} \left\{ p^* \exp\left[|c_1 + c_2 h| + |c_2 - (\mu(c_1^2 - 1)c_2 + c_1)h + \delta h| \right] \right. \\ & \left. + (1 - p^*)\exp\left[|c_1 + c_2 h| + |c_2 - (\mu(c_1^2 - 1)c_2 + c_1)h - \delta h| \right] \right\} \end{aligned} \quad (6')$$

Once again, this minimization can easily be performed by a computer, so that the functions $f_1(c_1,c_2)$ and $M_1(c_1,c_2)$, the minimizing value of μ for each state (c_1,c_2) can be taken as known.

Next we consider the process which is initiated at time $(N - k)h$ with the system in a general state (c_1,c_2), so that the process involves k decisions. We wish to determine the optimal first decision for the controller to make under these circumstances, and we denote the optimal value of μ for each state (c_1,c_2) by $M_k(c_1,c_2)$.

For any choice of the system parameter μ, the state of the system is transformed from the state (c_1,c_2) at time $(N - k)h$ into the state

$$(c_1 + c_2h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right]h + \delta h)$$

with probability p^* and into the state

$$(c_1 + c_2h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right]h - \delta h)$$

with probability $1 - p^*$. Consequently, once again using the principle of optimality, we obtain the recurrence relation

$$\begin{aligned} f_k(c_1,c_2) = \min_{m_1 \le \mu \le m_2} \Big\{ & |c_1|h + |c_2|h \\ & + p^* f_{k-1}(c_1 + c_2h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right]h + \delta h) \\ & + (1 - p^*) f_{k-1}(c_1 + c_2h, c_2 - \left[\mu(c_1^2 - 1)c_2 + c_1\right]h - \delta h) \Big\}, \end{aligned} \tag{7}$$

$k = 2,3,\ldots,N$. The term in brackets represents the cost during the first period from $(N - k)$ to $(N - k+1)$, and the second the minimal expected cost over the remaining $k - 1$ periods. As before in the deterministic case, we can determine computationally the desired functions $f_k(c_1,c_2)$ and $M_k(c_1,c_2)$, using the foregoing recursive relations.

7. Adaptive Control

In some circumstances, even less information than was assumed in the previous section will be available to the controller concerning external influences which may affect the behavior of the system being controlled. Provision may be made, though, for the controller to "learn" about the nature of these influences, as the process unfolds. It may then be able to improve its control decision-making capability in the course of time. In this sense the controller adapts itself to circumstances.

Observe that we are using the word in a quite precise sense. There is nothing mystical about the machine "thinking" or "creating" or "learning" in this restricted sense. That the human mind works in this way, or that the machine in any sense approximates the behavior of the human mind, can only be concluded on the basis of a rash evaluation of the possibilities of a digital computer or a brash contempt for the power of the human mind.

Let us return to our nonlinear system which is being disturbed by a random force. Let us now deprive the controller of the knowledge of the exact value of p. The controller still knows that $r_n = \pm \delta$, but the probability of each outcome is not known. Although this is an unpleasant situation, this controller is still much more fortunate than one that does not even know the form of the distributions of the variables r_n, or their degree of correlation. We shall not enter into a discussion of such matters here.

We can proceed with the design of an adaptive controller along the following lines. The state of the system will be characterized now not only by a position and a velocity, but also by a current estimate for p, which in the absence of further information we shall agree to regard as the precise value of the probability that $r_n = + \delta$

At any particular stage of the process when a control decision is made, not only does the system change state physically, but on the basis of the knowledge of the original physical state, the transformed physical state and the parameter value (μ) chosen, the controller can determine the sign of the unknown force for that stage. This may lead the controller to change its estimate of p. But how shall the estimate be changed?

Though there are many ways of answering this question, let us indicate one specific approach. Let us regard p itself as a random variable with an a priori probability density function $w(x)$, i.e.,

$$\text{Prob}\left\{a \le p \le a + \Delta\right\} = w(a)\Delta + o(\Delta^2). \tag{1}$$

As the initial estimate of p^*, a value we shall call p_1, we take the expected value of p,

$$p_1 = \int_0^1 xw(x)dx. \tag{2}$$

Upon observing that a positive disturbing force is exerted, $r = + \delta$, our new estimate of the probability density function for p will be given by

$$w_1(x) = \frac{xw(x)}{\int_0^1 xw(x)dx}. \tag{3}$$

Upon observing a negative disturbing force, $r = - \delta$, we shall change our estimate of the probability density function to

$$w_2(x) = \frac{(1 - x)w(x)}{\int_0^1 (1 - x)w(x)dx}. \tag{4}$$

Here we have adopted a Bayes approach, [15]. This is the procedure adopted in [6,7,9]. Consequently, after observing a positive disturbing force, the new estimate of p^* itself is

$$\frac{\int_0^1 x^2 w(x)dx}{\int_0^1 xw(x)dx}, \tag{5}$$

and after observing a negative disturbing force, the new estimate of p* is

$$\frac{\int_0^1 x(1-x)w(x)dx}{\int_0^1 (1-x)w(x)dx}. \tag{6}$$

By way of summary we may note that if the a priori choice of probability density function is w(x), then after m positive and n negative forces have been observed the new estimate of the density function is

$$\frac{x^m(1-x)^n w(x)}{\int_0^1 x^m(1-x)^n w(x)dx},$$

and $p_{m,n}$, the new estimate of p* itself, is

$$p_{m,n} = \frac{\int_0^1 x^{m+1}(1-x)^n w(x)dx}{\int_0^1 x^m(1-x)^n w(x)dx}. \tag{7}$$

The controller is to act as if this estimate is the exact value of p*. This should be recognized as an assumption of our analysis.

The integrals in equation (7) simplify if w(x) is the density function for a beta distribution,

$$w(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \quad a,b > 0, \tag{8}$$

where B(a,b) is the beta function. Great flexibility in the shape of the curve of w(x) can be achieved by selecting the parameters a and b appropriately. For this choice of w(x) we have

$$p_{m,n} = \frac{\int_0^1 x^{m+a}(1-x)^{n+b-1}dx}{\int_0^1 x^{m+a-1}(1-x)^{n+b-1}dx} \tag{9}$$

$$= \frac{B(m+a+1,\ n+b)}{B(m+a,\ n+b)}$$

$$= \frac{(m+a)}{(m+a)+(n+b)}.$$

The parameters a and b play the roles of the a priori numbers of positive and negative forces observed. If the sum is small, not much weight is given to the initial estimate; if it is large, many periods of the process are required before the estimate of p* can be significantly changed.

We now take up the problem of determining the optimal decisions for the controller to make. First the function w(x) is chosen and it is then held fixed. Then we define the sequence of functions.

$f_k(c_1,c_2;m,n)$ = the expected cost of control over the last k stages of a process in which initially the system is in physical state (c_1,c_2) and m positive and n negative forces have been observed, using an optimal control policy. (10)

In taking expected values we shall assume that current estimates of distributions are the true distributions, consistent with our previous practice.

For the k-stage process we find (11)

$$f_k(c_1,c_2;m,n) = |c_1|h + |c_2|h + \underset{m_1\le\mu\le m_2}{\text{Min}} \Big\{ p_{m,n} f_{k-1}(c_1+c_2h,\ c_2-\big[\mu(c_1^2-1)c_2 + c_1\big]h+\delta h;\ m+1,n) + (1-p_{m,n}) f_{k-1}(c_1+c_2h,\ c_2-\big[\mu(c_1^2-1)c_2 + c_1\big]h-\delta h;\ m,n+1) \Big\},$$

for k= 2,3,...,N, and for the one-stage process

$$f_1(c_1,c_2;m,n) = |c_1|h + |c_2|h + \underset{m_1\le\mu\le m_2}{\text{Min}} \Big\{ p_{m,n}\exp(|c_1+c_2h| + |c_2-\big[\mu(c_1^2-1)c_2 + c_1\big]h+\delta h|) + (1-p_{m,n})\exp(|c_1+c_2h| + |c_2-\big[\mu(c_1^2-1)c_2 + c_1\big]h-\delta h|) \Big\}. \tag{12}$$

As before, the functions f_k and the decision functions $M_k(c_1,c_2;m,n)$ are to be determined recursively.

The numerical resolution of equations (11) and (12) presents some difficulties, however, in that sequences of functions of four arguments are to be determined. Several methods present themselves for consideration, however. In particular, we note that when m + n is large we can assert with some confidence that m/(m + n) is a good estimate for p*. The decisions called for in the solution of the stochastic control process discussed in §6 should provide nearly optimal control decisions. Some advantage may be gained by considering an infinite stage process as in §5, which has the effect of eliminating one subscript. These, and other matters, will be discussed in a forthcoming thesis by M. Aoki, [11].

Notice that the adaptive controller discussed in this section does no "research" regarding the random variables. It merely observes the history of the process to date and combines this with its a priori knowledge in a way which is specified at the beginning of the process, in order to arrive at its current control decision. How the controller will act as the result of any particular observed history is fully specified initially. More sophisticated controllers will be designed to look for correlations, provide for non-stationarities in the unknown process, and so on. Much remains to be done along these lines.

8. An Extension

Let us now turn our attention to the case in which the system can be controlled both by modifying the system parameter μ and also by exerting a control force, g_k, at time kh, where $k = 0,1,2,\ldots,N-1$. The equation for the state of the system becomes

$$x_{n+1} = x_n + v_n h, \quad x_0 = c_1, \tag{1}$$

$$v_{n+1} = v_n - \left[\mu_n(x_n^2 - 1)v_n + x_n\right]h + r_n h + g_n h, \quad v_0 = c_2,$$

where now both μ_n and g_n are at our disposal. As before, μ_n is limited to lie within the region

$$m_1 \le \mu_n \le m_2, \quad n = 0, 1, \ldots, N-1, \tag{2}$$

and we shall also assume that the control force g_n is constrained by the inequality

$$|g_n| \le G, \quad n = 0,1,2,\ldots N-1. \tag{3}$$

Upon introduction of the functions $f_k(x,v;m,n)$,

$$f_k(c_1,c_2;m,n) = \min_{\substack{m_1\le\mu_j\le m_2 \\ |g_j|\le G}} E\left[\sum_{j=N-k}^{N-1}(|x_j|+|v_j|)h + \exp(|x_N|+|v_N|)\right], \tag{4}$$

for $k = 1,2,\ldots,N$, we find that the relevant functional equations are

$$f_k(c_1,c_2;m,n) = |c_1|h + |c_2|h + \min_{\substack{m_1\le\mu\le m_2 \\ |g|\le G}}\Big\{p_{m,n} f_{k-1}(c_1+c_2h, c_2-\left[\mu(c_1^2-1)c_2+c_1\right]h + (\delta+g)h;m+1,n) + (1-p_{m,n})f_{k-1}(c_1+c_2h, c_2-\left[\mu(c_1^2-1)c_2+c_1\right]h + (-\delta+g)h;m,n+1)\Big\}, \tag{5}$$

$k = 2,3,\ldots,$

and

$$f_1(c_1,c_2;m,n) = |c_1|h + |c_2|h + \min_{\substack{m_1\le\mu\le m_2 \\ |g|\le G}}\Big\{p_{m,n}\exp(|c_1+c_2h|+|c_2-\left[\mu(c_1^2-1)c_2+c_1\right]h + (\delta+g)h|) + (1-p_{m,n})\exp(|c_1+c_2h|+|c_2-\left[\mu(c_1^2-1)c_2+c_1\right]h + (-\delta+g)h|)\Big\}. \tag{6}$$

This case differs mathematically from the previous ones only in that the minimizations are now over a two-dimensional region rather than a one-dimensional region.

9. Discussion

In the earlier sections of this paper we have sketched a treatment of an adaptive control process from the dynamic programming viewpoint. Much remains to be done at various levels in the treatment of these fascinating control processes.

At the conceptual level, for example, models involving other types of uncertainties on the part of the controller, mentioned in §1, have yet to be constructed. One of the principal difficulties occurs in describing the state of knowledge of the controller and how this changes as new information is added. Furthermore, so much information may become available that a way must be found to summarize it succinctly without impairing the decision-making capability to a marked degree. In this connection, see the discussion of sufficient statistics in [14].

Insofar as the mathematical analysis itself is concerned, many perplexing problems arise, as, for example, questions concerning the convergence of discrete adaptive processes to continuous processes and the very formulation of adaptive processes of continuous type.

Lastly, as we have already indicated, for the more realistic processes involving more state variables, the computational solutions present special problems of their own, all of which must be carefully investigated.

Another whole problem area beyond those already mentioned is encompassed by the actual construction of optimal adaptive controllers. Challenging problems arise in trying to pursue a straight and narrow path between the complexity of exact solutions and the fallibility of approximations.

References

1. Stoker, J. J., Nonlinear Vibrations in Mechanical and Electrical Systems, Interscience Publishers, Inc., New York, 1950.

2. Bellman, R., 'On the application of the theory of dynamic programming to the study of control processes,' Proc. Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, New York, 1957, pp. 199-213.

3. Bellman, R., 'Dynamic programming and stochastic control processes,' Information and Control, vol. 1, 1958, pp. 228-239.

4. Bellman, R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

5. Bellman, R., and R. Kalaba, 'On the role of dynamic programming in statistical communication theory,' IRE Trans. on Information Theory, vol. IT-3, 1957, pp. 197-203.

6. Bellman, R., and R. Kalaba, 'On communication processes involving learning and random duration,' 1958 IRE National Convention Record, part 4, July 1958, pp. 16-21.

7. Bellman, R., and R. Kalaba, Dynamic programming and adaptive processes--I: Mathematical Foundation, The RAND Corporation, Paper P-1416, July 3, 1958.

8. Bellman, R., and R. Kalaba, 'On the principle of invariant imbedding and propagation through inhomogeneous media,' Proc. Nat. Acad. Sci. USA, vol. 42, 1956, pp. 629-632.

9. Bellman, R., 'A problem in sequential design of experiments,' Sankhya, vol. 16, 1956, pp. 221-229.

10. Aseltine, J. A., A. R. Mancini, and C. W. Sarture, 'A Survey of adaptive control systems,' IRE Trans. on Automatic Control, PGAC-6, Dec. 1958, pp. 102-108.

11. Aoki, M., Ph.D. Thesis, University of California at Los Angeles, to appear.

12. Freimer, M., Ph.D. Thesis, Harvard University, to appear.

13. Bellman, R., and R. Kalaba, On k-th best policies, The RAND Corporation, Paper P-1417, July 1958.

14. Mood, A. M., Introduction to the Theory of Statistics, McGraw-Hill Book Company, Inc., New York, 1950.

15. Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey, 1951

16. Courant, R., and D. Hilbert, Methods of Mathematical Physics, vol. 1, Interscience Publishers, Inc., New York, 1953.

A DYNAMIC PROGRAMMING APPROACH TO ADAPTIVE CONTROL PROCESSES*

Marshall Freimer

Staff Member
Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Massachusetts

Summary

In many multi-stage decision processes we face the problem of dealing with random variables whose distributions are initially imperfectly known, but which become known with increasing accuracy as the process continues. In this paper we shall show how Dynamic Programming[1] may be used to treat a class of such problems, which are currently called adaptive processes.

After discussing the general theory, we shall illustrate the techniques by a specific example. For this example we derive simple computational algorithms, which are typical of those obtained for the whole class of problems under consideration.

1. Introduction

A large class of control problems concerns systems which can be described at any time t by a real s-dimensional state vector, x(t), obeying a linear differential equation

$$\dot{x}(t) + A(t)x(t) = y(t) \quad . \tag{1}$$

Here y(t) is a real s-dimensional control vector, and A(t) is a known s x s matrix function of time. We are given the initial state

$$x(0) = c \quad , \tag{2}$$

and are asked to choose y(t) so as to minimize the error functional

$$G_T(y) = E\left\{\int_0^T g(x(t), y(t), r(t), t)dt\right\}, \tag{3}$$

where r(t) is an s-dimensional stochastic process, and g is a given function.

2. Discrete Formulation

This problem becomes particularly interesting when the statistical properties of the stochastic process r(t) are incompletely known. For this purpose, the continuous version described above is rather cumbersome. Instead, we shall suppose that time is measured discretely, and that at any time n we have x(n), y(n), and r(n), analogous to x(t), y(t), and r(t) above. The control equation may now be written as

$$x(n) = A(n)x(n-1) + y(n) \quad , \tag{4}$$

the initial condition as

$$x(0) = c \quad , \tag{5}$$

and the error functional to be minimized as

$$G_N(y) = \sum_{n=1}^{N} \left\{E\ g(x(n), y(n), r(n), n)\right\} . \tag{6}$$

In many processes time is measured most naturally in discrete periods, so that the continuous process described by equation (1) is only a mathematical idealization. For such a descrete process, equation (4) is more suitable.

A further advantage to be found in using the discrete formulation is that the solution appears as a set of recurrence formulas. These are ideally suited for automatic programming by a digital computer.

3. The Stochastic Process r(n)

We must now clarify what we mean by a stochastic process with imperfectly known distribution. There are several ways to do this, but we shall consider just one.

We shall assume that r(n) is a stationary Markov chain, of some order ν. (The simplest case is $\nu = 0$, where the r(n) are independent). We suppose that we know the transition probabilities in terms of certain unknown but fixed parameters, and that we have an a priori distribution for the values of parameters. As the process proceeds, we observe r(1), r(2), . . . , r(m) and summa-

*The work reported here was performed at Lincoln Laboratory, a technical center operated by Massachusetts Institute of Technology with the joint support of the Army, Navy and Air Force.

rize them by a sufficient statistic S(m). Using S(m) and the a priori distribution, we compute an a posteriori distribution for the unknown parameters, and this gives us an up-to-date distribution to use for r(m+1).

We can show that the a posteriori distributions for the unknown parameters will converge, with probability one, to the true values of these parameters. At the same time, the up-to-date distribution for the r's will converge to the true distribution.

The following examples should serve to clarify the concept of a sufficient statistic.

Example 1: ($\nu = 0$). Suppose that r(n) is a sequence of independent, identically distributed Gaussian random variables, with mean 0 and unknown variance σ^2. Let us also assume that the priori distribution is a γ-distribution on $1/\sigma^2$, $\gamma(1/\sigma^2; a, b)$, so that the density function for σ is

$$\mu(\sigma) = \frac{-2}{\sigma^3} \frac{a^b}{\Gamma(b)} \frac{1}{\sigma^{2(b-1)}} \exp\left(-\frac{a}{\sigma^2}\right), \qquad \sigma \geq 0, \quad (7)$$

where a and b are given positive numbers.[2] Then

$$S(m) = \sum_{n=1}^{m} r(n)^2 \quad (8)$$

is a suitable sufficient statistic, since it enables us to compute the a posteriori density

$$\mu_m(\sigma) = \frac{-2}{\sigma^3} \frac{\left[a + \frac{1}{2}S(m)\right]^{b+\frac{1}{2}m}}{\Gamma(b+\frac{1}{2}m)} \cdot \frac{1}{\sigma^{2(b+\frac{1}{2}m-1)}} \exp\left(-\frac{a+\frac{1}{2}S(m)}{\sigma^2}\right), \quad (9)$$

which gives the γ-distribution $\gamma(\frac{1}{\sigma^2}; a+\frac{1}{2}S(m), b+\frac{1}{2}m)$ for $\frac{1}{\sigma^2}$. It is this reproductive property which led us to use an a priori γ-distribution.

To find the distribution of r(m+1) we multiply $\mu_m(\sigma)$ by the Gaussian distribution and integrate over σ. Doing this, we obtain the up-to-date density function

$$\lambda_m(r) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(b+\frac{1}{2}(m+1))}{\Gamma(b+\frac{1}{2}m)} \frac{\left[a+\frac{1}{2}S(m)\right]^{b+\frac{1}{2}m}}{\left[a+\frac{1}{2}S(m)+\frac{1}{2}r^2\right]^{b+\frac{1}{2}(m+1)}} \quad (10)$$

for r(m+1). If we make the substitution

$$t^2 = r^2 \frac{2b+m}{2a+S(m)}, \quad (10a)$$

we find that this is Student's t-distribution with 2b + m degrees of freedom.

It is also important to be able to calculate S(m+1) from S(m) and r(m+1). This is readily done using the formula

$$S(m+1) = S(m) + r(m+1). \quad (11)$$

Example 2: ($\nu = 1$). Suppose that r(n) is a first order Markov chain of 0-1 random variables. If r(n-1) = 1, let θ_1 and $1-\theta_1$ be the probabilities of r(n) = 1 and r(n) = 0, while if r(n-1) = 0 let θ_0 and $1-\theta_0$ be the respective probabilities of 1 and 0. We take θ_0 and θ_1 as having independent β-distributions, so that the a priori density is

$$\mu(\theta_0, \theta_1) = \frac{\theta_0^{a-1}(1-\theta_0)^{b-1}}{B(a,b)} \frac{\theta_1^{c-1}(1-\theta_1)^{d-1}}{B(c,d)} \quad (12)$$

where a, b, c, and d are arbitrary positive numbers.

Let m(i, j) be the number of pairs $\{r(n-1)=i, r(n)=j\}$ observed for n = 1, 2, . . . , m. Then as a sufficient statistic we can take the vector

$$S(m) = (m(0,0), m(0,1), m(1,0), m(1,1)). \quad (13)$$

The a posteriori distribution is

$$\mu_m(\theta_0, \theta_1) = \frac{\theta_0^{a+m(0,1)-1}(1-\theta_0)^{b+m(0,0)-1}}{B(a+m(0,1), b+m(0,0))} \cdot \frac{\theta_1^{c+m(1,1)-1}(1-\theta_1)^{d+m(1,0)-1}}{B(c+m(1,1), d+m(1,0))} \quad (14)$$

In this simple case we can compute the expected values of θ_0 and θ_1 with respect to $\mu_m(\theta_0, \theta_1)$,

and use the appropriate one as the probability of $r(m+1)=1$. We find that these expected values are

$$E\{\theta_0\} = \frac{a + m(0,1)}{a + b + m(0)}$$

$$E\{\theta_1\} = \frac{c + m(1,1)}{c + d + m(1)} \quad , \qquad (15)$$

where $m(0) = m(0,0) + m(0,1)$ is the number of $r(n)$ which equal 0, $n = 1, 2, \ldots, m$, and similarly for $m(1) = m(1,0) + m(1,1)$.

In order to obtain $S(m+1)$ from $S(m)$ we need both $r(m+1)$ and $r(m)$. But we already needed $r(m)$ to determine whether θ_0 or θ_1 should be used for $r(m+1)$, so this is no additional burden.

Thus we should use

$$S(m) = (m(0,0), m(0,1), m(1,0), m(1,1), r(m)). \qquad (13')$$

The formula for $S(m+1)$ in terms of $S(m)$ and $r(m+1)$ is then readily obtained.

4. The Functional Equation

At time m knowledge of the state variable $z = (x(m-1), S(m-1))$ completely specifies the current status of the problem. If $y(m)$ is our choice of control vector, then the new state variable is given by (16)

$$T_{y(m)}(z) = (x(m), S(m)) \qquad (17)$$

where $x(m)$ is computed by equation (4), and $S(m)$ is computed by a suitable equation (such as (11)).

Note that at time m z is deterministic, but that $T_{y(m)}(z)$ is a random variable.

Knowing $y(m)$ and z we can also compute the expected value of the immediate error,

$$g(m) = E\{g(x(m), y(m), r(m), m)\} \quad , \qquad (18)$$

where the expectation is with respect to the up-to-date distribution for $r(m)$ (e.g. (10) or (15)).

Thus, knowledge of z and choice of $y(m)$ lead to an expected immediate error of $g(m)$, and a transformed state variable $T_{y(m)}(z)$.

Now let $f_m(z)$ be the minimum expected loss achievable, starting at time m and with state variable z. We may write

$$f_m(z) = \min_{y(m), \ldots, y(N)} \sum_{n=m}^{N} E\{g(x(n), y(n) r(n), n)\} . \qquad (19)$$

We shall apply the Principle of Optimality[3] of Dynamic Programming to obtain a functional equation relating f_m with f_{m+1}. Let us rewrite the right side of equation (19) in the form

$$\min_{y(m)} \left[E\, g(x(m), y(m), r(m), m) + \min_{y(m+1), \ldots, y(N)} \sum_{n=m+1}^{N} E\{g(x(n), y(n), r(n), n)\} \right] ,$$

and observe that the values of $y(m+1), \ldots, y(N)$ which furnished the minimum in equation (19) also minimize the second term here. This second term thus becomes $f_{m+1}(T_{y(m)}(z))$, applying definition (19) at time $m+1$. If we substitute this into equation (19) we obtain the functional equation

$$f_m(z) = \min_{y(m)} E\{g(x(m), y(m), r(m), m) + f_{m+1}(T_{y(m)}(z))\} \quad . \qquad (20)$$

The problem requires us to compute $f_1(c)$, where we have written c instead of $(c, S(0))$ since $S(0)$ contains no information at all. We also need the minimizing choice of $y(1), y(2), \ldots, y(N)$, known as the optimal policy, but these will arise naturally in the calculations.

The problem is solved by noting that

$$f_N(z) = \min_{y(N)} E\{g(x(N), y(N), r(N), N)\} \quad , \qquad (21)$$

for $z = (x(N-1), S(N-1))$, can be found by differential calculus. We then apply the functional equation (20) repeatedly, obtaining $f_{N-1}, f_{N-2}, \ldots, f_1$.

5. Quadratic Error Functional

The solution just obtained involves the recursive computation of a function of $s + d$ real variables, where d is the dimension of each $S(n)$. This will generally require far greater computational capacity than even the largest digital computers now possess.

In order to overcome this difficulty we shall specialize $g(x(n), y(n), r(n), n)$ to be a quadratic function of the components of $x(n)$, $y(n)$, and $r(n)$. We can then show that $f_n(z)$ is a quadratic function of the components of $x(n-1)$, with coefficients depending on $S(n-1)$. Simple recurrence relations for these coefficients can be obtained from equation (20), so that the problem of recursive computation has been reduced from one of $s + d$ to one of d dimensions.

6. An Illustrative Example

To illustrate the techniques described in the preceeding sections, we shall consider a one-dimensional tracking problem. This may be thought of as one coordinate of a two- or three-dimensional problem, since all effects can be seen to be additive.

Suppose that at time n we are located at the point x(n), while our quarry is at r(n). Our motion is controlled by the equation

$$x(n+1) = x(n) + y(n+1) \tag{22}$$

where y(n) is our control variable. His motion is controlled by

$$r(n+1) = r(n) + u(n+1) \tag{23}$$

where the u's are independent, identically distributed Gaussian random variables, with mean 0 and unknown variance σ^2. The a priori distribution for σ is $(\frac{1}{\sigma^2}; a, b)$, whose density function is given by equation (7). We shall certainly need

$$S_1(m) = \sum_{n=1}^{m} u(n)^2 \tag{24}$$

as part of our sufficient statistic for determining the distribution of r(m+1), since we use this for the variance. But we must also find the mean, which is given by

$$S_2(m) = r(m) = \sum_{n=1}^{m} u(n) \quad . \tag{25}$$

We can then take

$$S(m) = (S_1(m),\ S_2(m)) \quad . \tag{26}$$

The a posteriori distribution of r(m+1) is then given by its density function

$$\lambda_m(r) = \frac{1}{2\pi} \frac{\Gamma(b+\frac{1}{2}(m+1))}{\Gamma(b+\frac{1}{2}m)} \frac{\left[a+\frac{1}{2}S_1(m)\right]^{b+\frac{1}{2}m}}{\left[a+\frac{1}{2}S_1(m)+\frac{1}{2}(r-S_2(m))^2\right]^{b+\frac{1}{2}(m+1)}} . \tag{27}$$

Finally, the error function is given as

$$g(x(n), y(n), r(n), n) = a(n)y(n)^2 + b(n)\left[x(n)-u(n)\right]^2, \tag{28}$$

where a(n) and b(n) are known, positive functions of time. The terms on the right in equation (28) may be interpreted as an energy cost for moving us about, and a cost for inexact tracking.

Our problem is to choose y(1), y(2), . . . , y(N) so as to minimize the error functional

$$G_N(y) = \sum_{n=1}^{N} E\left\{g(x(n), y(n), r(n), n)\right\}, \tag{29}$$

given the initial conditions

$$x(0) = c$$
$$u(0) = 0 \quad . \tag{30}$$

As our state variable at time m we take

$$z = (x(m-1), S(m-1)) \quad . \tag{31}$$

If we choose y(m) then we get the transformed state variable

$$T_{y(m)}(z) = (x(m), S(m)) \ , \tag{32}$$

where x(m) is given by equation (22), and

$$\begin{aligned} S_1(m) &= S_1(m-1) + u(m)^2 \\ S_2(m) &= S_2(m-1) + u(m) \quad . \end{aligned} \tag{33}$$

We can now define $f_m(z)$, the minimum expected loss starting at time m and state variable z, as in equation (19). We assert that this function is a quadratic function of x(m-1), say

$$f_m(z) = \alpha x(m-1)^2 + \beta x(m-1) + \gamma \tag{34}$$

where the coefficients α, β, and γ are functions of m and S(m-1). If we substitute this expression for $f_m(z)$, and the corresponding expression for $f_{m+1}(T_{y(m)} z)$, into equation (20) we get

$$\begin{aligned} &\alpha(m, S(m-1))x(m-1)^2 + \beta(m, S(m-1))x(m-1) + \gamma(m, S(m-1)) \\ &= \min_{y(m)} E\Big\{a(m)y(m)^2 + b(m)\left[x(m)-r(m)\right]^2 \\ &\quad + \alpha(m+1, S(m))\gamma(m)^2 + \beta(m+1, S(m))x(m) + \gamma(m+1, S(m))\Big\} \end{aligned} \quad . \tag{35}$$

It is easier to deal with this equation if we replace y(m) by x(m)-x(m-1), and choose x(m) so as to achieve the minimum. The minimizing choice of x(m), found by setting the derivative equal to 0, is

$$x(m) = \frac{a(m)x(m-1) + b(m)E\left\{r(m)\right\} - \frac{1}{2}E\left\{\beta(m+1, S(m))\right\}}{a(m) + b(m) + E\left\{\alpha(m+1, S(m))\right\}} \quad . \tag{36}$$

The expectations in equation (36) are to be taken with respect to the a posteriori distribution for r(m), given by equation (27). We also note that they are abbreviations, e.g.,

$$E\{\alpha(m+1, S(m))\} = E\{\alpha(m+1, S_1(m-1)+r(m)^2, S_2(m-1)+r(m))\} \tag{37}$$

If we substitute expression (36) into equation (35), and equate the coefficients of like powers of x(m-1), we obtain

$$\alpha(m, S(m-1)) = \frac{a(m)[b(m)+E\{\alpha(m+1, S(m))\}]}{a(m)+b(m)+E\{\alpha(m+1, S(m))\}}, \tag{38}$$

$$\beta(m, S(m-1)) = \frac{a(m)[E\{\beta(m+1, S(m))\} - 2b(m)E\{r(m)\}]}{a(m)+b(m)+E\{\alpha(m+1, S(m))\}} \tag{39}$$

and

$$\gamma(m, S(m-1)) = E\{\gamma(m+1, S(m))\} + b(m)E\{r(m)^2\} - \frac{[b(m)E\{r(m)\} - \frac{1}{2}E\{\beta(m+1, S(m))\}]^2}{a(m)+b(m)+E\{\alpha(m+1, S(m))\}} . \tag{40}$$

Using equation (21) we can also find that

$$\alpha(N, S(N-1)) = \frac{a(N)b(N)}{a(N)+b(N)} , \tag{38'}$$

$$\beta(N, S(N-1)) = \frac{2a(N)b(N)E\{r(N)\}}{a(N)+b(N)} , \tag{39'}$$

and

$$\gamma(N, S(N-1)) = b(N)E\{r(N)^2\} - \frac{b(N)^2E\{r(N)\}^2}{a(N)+b(N)} . \tag{40'}$$

We can observe several important facts from these equations. First, from equation (38') we see that $\alpha(N, S(N-1))$ does not depend on S(N-1). Working backwards, by means of equation (38), we then find that for any m, $\alpha(m, S(m-1))$ does not depend on S(m-1). We shall therefore write $\alpha(m)$ in place of $\alpha(m, S(m-1))$. Thus equation (38) becomes

$$\alpha(m) = \frac{a(m)[b(m) + \alpha(m+1)]}{a(m) + b(m) + \alpha(m+1)} , \tag{41}$$

and similarly for equation (38').

Secondly, from equations (36), (38), (38'), (39) and (39') we see that we do not need the γ's. We thus can dispense with equations (40) and (40'), again reducing our workload considerably.

Finally, computation of the β's from equations (39) and (39') is not as difficult as it might appear at first. If we study these equations we find that the expectations involved are all linear in the r's, and that the apparently compound expectations reduce to simple ones according to the rule that the expectation of a conditional expectation is just the unconditional expectation. Thus, in equation (39') we can put

$$E\{r(N)\} = S_2(N-1) , \tag{42}$$

since $S_2(m)$ is the mean of the distribution $\lambda_m(r)$ of r(m+1), (equation (27)). Then

$$\beta(N, S(N-1)) = -\frac{2a(N)b(N)S_2(N-1)}{a(N) + b(N)} . \tag{43}$$

Now, in using equation (39) at time m = N-1 we can put

$$E\{r(N-1)\} = S_2(N-2) \tag{44}$$

and

$$E\{\beta(N, S(N-1))\} = -\frac{2a(N)b(N)E\{S_2(N-1)\}}{a(N)+b(N)} = -\frac{2a(N)b(N)S_2(N-2)}{a(N)+b(N)} , \tag{45}$$

since $S_2(N-1) = r(N-1)$ by equation (25). This may be written as

$$E\{\beta(N, S(N-1))\} = \beta(N, S_2(N-2)). \tag{46}$$

If we work our way backwards in time we find that at any time m we have

$$E\{r(m)\} = S_2(m-1) \tag{47}$$

and

$$E\{\beta(m+1, S(m))\} = \beta(m+1, S_2(m-1)) , \tag{48}$$

where β is no longer a function of $S_1(m-1)$. Thus equation (39) becomes

$$\beta(m, S_2(m-1)) = \frac{a(m)[\beta(m+1, S_2(m-1)) - 2b(m)S_2(m-1)]}{a(m)+b(m)+\alpha(m+1)} . \tag{49}$$

Similarly, equation (36) can now be written as

$$x(m) = \frac{a(m)x(m-1) + b(m)S_2(m-1) - \frac{1}{2}\beta(m+1, S_2(m-1))}{a(m) + b(m) + \alpha(m+1)} . \tag{50}$$

Equations (38'), (41), (43), (49), and (50) give us the simple computational algorithm promised in the summary. Examining these equations we find that we do not need to compute any distributions at all. This is a basic property of the Dynamic Programming solution of any linear-control, quadratic-error process.

References

1. R. Bellman, Dynamic Programming, Princeton University Press, 1957.

2. H. Raiffa and R. Schlaifer, Informal notes, Harvard Business School, (Unpublished).

3. R. Bellman, op. cit., p. 83.

ON THE OPTIMUM SYNTHESIS OF MULTIPOLE CONTROL SYSTEMS IN THE WEINER SENSE *

H. C. Hsieh and C. T. Leondes
University of California, Los Angeles

ABSTRACT

This paper is concerned with obtaining the optimum system in the Weiner sense for the multipole system shown in Figure 1. Earlier literature[1] has shown how to obtain the mean-square value of the error when the multipole system transfer function has been specified, but thus far no published work has shown how to solve the synthesis problem, in general, for this case. The principal reason that this problem has appeared to be impossible of analytic solution thus far for cross correlation between the inputs is based on the fact that the usual variational approach results in a set of untractable simultaneous integral equations involving many complicated cross products of the desired weighting functions and the variational functions.

The synthesis problem for the system of Figure 1 is first solved for the case in which there is no correlation between the inputs to the various terminals. The result for the optimum weighting functions in this case is presented in equation (24), and the resultant mean-squared value of the error is shown in equation (25).

Following this, the far more complicated case of the synthesis problem when the inputs to all the various terminals are correlated is considered. In this case, a rather unique technique is utilized to avoid the difficulties inherent in the use of the usual variational techniques. Through the technique utilized in this paper, the usual set of untractable simultaneous integral equations is completely avoided, and instead a set of ordinary algebraic equations results. The set of equations for this case is shown in equation (64), and in matrix form in equation (65). The resultant solution for the optimum physically realizable transfer functions is shown in equation (77). It is also shown, as a check, that the solution for the case of correlated inputs reduces to the solution obtained for the case of uncorrelated inputs.

The paper then concludes with an illustrative example for the more complicated case of correlated inputs. The possibilities of applications of the results of this paper to such fields as the guidance and control of astronautical vehicles, military fire control systems, bombing navigation systems, process control systems, automatic milling machines, air traffic control, nuclear reactor control, etc., are fairly evident.

* This work was done under contract AF49(638)-438 with the Air Force Office of Scientific Research.

Optimum Multipole Control Systems with Uncorrelated Stationary Signals and Noises Between Terminals.

The systems under consideration are linear. For a linear system it is possible to superimpose the effects produced by any number of inputs. The response characteristics of linear systems can be represented by the weighting function or the transfer function. The result of employing either of these two methods are equivalent, each having its own merits for particular classes of problems.

The weighting function $W(t,\tau)$ is the unit impulse response of the system. $W(t,\tau)$ depends upon the observing time t and also upon the time at which the impulse is applied. For a system describable in terms of ordinary linear differential equations with constant coefficients, the unit impulse response depends only upon the interval between application of the impulse and observation of the output. Therefore, for such a system

$$w(t,\tau) = W(t-\tau) \tag{1}$$

As is well known, the usefulness of the weighting function lies in the fact that it permits a convenient representation of system output to be made in terms of the corresponding input. Thus the output, y(t), of a system can be represented in terms of the input, x(t), through the equation

$$y(t) = \int_{-\infty}^{t} W(t-\tau)\, x(\tau)\, d\tau \tag{2}$$

By a change of variable, this relationship may also be written

$$y(t) = \int_{0}^{\infty} W(\tau)\, x(t-\tau)\, d\tau \tag{3}$$

If the input to a system is sinusoidal, and can be expressed by complex quantity e^{jwt}, then from equation (3) we have

$$y(t) = \int_{0}^{\infty} W(\tau) e^{j\omega(t-\tau)} d\tau = e^{j\omega t} \int_{0}^{\infty} W(\tau) e^{-j\omega\tau} d\tau \tag{4}$$

Thus it is seen that the output of the system differs from the input only by a constant complex factor

$$Y(jw) = \int_{0}^{\infty} W(\tau)\, e^{-j\omega\tau}\, d\tau \tag{5A}$$

This quantity Y(jw) is called the frequency-response function. If we take the Laplace transform of W(t) rather than the Fourier transform, we shall have

$$Y(p) = \int_0^\infty W(t)\, e^{-pt}\, dt \qquad (5B)$$

where p is a complex constant. Here Y(p) is called the transfer function of the system.

The multipole control systems under consideration can be described by n sets of linear time-invariant input-output relations as

$$y_j(t) = \sum_{k=1}^{n} \int_0^\infty W_{jk}(\tau)\,[s_k(t-\tau) + n_k(t-\tau)]\, d\tau \qquad (6)$$
$$j = 1, 2, \ldots\ldots m$$

where s(t) is the signal or the desired input, and n(t) is the noise or the unwanted input. In matrix form, the system configuration is shown in Figure 1.

Now, if we let $D_{jk}(\tau)$ be the weighting function of an ideal system that accomplishes perfectly the desired task, then the desired output $z_j(t)$ can be seen to be

$$z_j(t) = \sum_{k=1}^{n} \int_0^\infty D_{jk}(\tau)\, s_k(t-\tau)\, d\tau \qquad (7)^*$$
$$j = 1, 2, \ldots\ldots m$$

Thus the difference between the actual and desired system outputs may be expressed as

$$\epsilon_j(t) = y_j(t) - z_j(t) = \sum_{k=1}^{n} \int_0^\infty W_{jk}(\tau)\,[s_k(t-\tau) + n_k(t-\tau)]\, d\tau$$
$$- \sum_{k=1}^{n} \int_0^\infty D_{jk}(\tau)\, s_k(t-\tau)\, d\tau \qquad (8)$$

A diagram of this general problem is given in Figure 2 where, for simplicity, the usual matrix notation is adopted. The part of the diagram below the dotted line represents the actual system, while the part above the dotted line represents the hypothetical ideal system together with comparator for generating the error signal.

After the error has been generated, our problem now is how to choose the weighting functions of the actual system so as to minimize some suitable function of the error (performance index). This performance index can, of course, take forms other than the least mean-squared value $\overline{\epsilon^2}$; for example, the quantity $|\overline{\epsilon}|$. However, the mean-squared error criterion in addition to providing a good indication of the quality of the system is very convenient to analyze mathematically, and therefore is most commonly used.

To formulate the mean-squared error, we first replace t and τ in equation (8) by (t_1, τ_1) and (t_2, τ_2), respectively, to obtain the two equations

$$\epsilon_j(t_1) = \sum_{k=1}^{n} \int_0^\infty W_{jk}(\tau_1)\,[s_k(t_1-\tau_1) + n_k(t_1-\tau_1)]\, d\tau_1$$
$$- \sum_{k=1}^{n} \int_0^\infty D_{jk}(\tau_1)\, s_k(t_1-\tau_1)\, d\tau_1$$

$$\epsilon_j(t_2) = \sum_{k'=1}^{n} \int_0^\infty W_{jk'}(\tau_2)\,[s_{k'}(t_2-\tau_2) + n_{k'}(t_2-\tau_2)]\, d\tau_2 \qquad (9)$$
$$- \sum_{k'=1}^{n} \int_0^\infty D_{jk'}(\tau_2)\, s_{k'}(t_2-\tau_2)\, d\tau_2$$
$$k, k' = 1, 2, \ldots\ldots n$$

Now we may compute the mathematical expectation of the product on both sides of these two equations

$$\overline{\epsilon_j(t_1)\epsilon_j(t_2)} =$$
$$\sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty W_{jk}(\tau_1)\, d\tau_1 \int_0^\infty \overline{[s_k(t_1-\tau_1) - n_k(t_1-\tau_1)]\,[s_{k'}(t_2-\tau_2) + n_{k'}(t_2-\tau_2)]}\, W_{jk'}(\tau_2)\, d\tau_2$$
$$- \sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty D_{jk}(\tau_1)\, d\tau_1 \int_0^\infty \overline{s_k(t_1-\tau_1)\,[s_{k'}(t_2-\tau_2) + n_{k'}(t_2-\tau_2)]}\, W_{jk'}\tau_2\, d\tau_2 \qquad (10)$$
$$- \sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty W_{jk}(\tau_1)\, d\tau_1 \int_0^\infty \overline{[s_k(t_1-\tau_1) + n_k(t_1-\tau_1)]\, s_{k'}(t_2-\tau_2)}\, D_{jk'}(\tau_2)\, d\tau_2$$
$$+ \sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty D_{jk}(\tau_1)\, d\tau_1 \int_0^\infty \overline{s_k(t_1-\tau_1)\, s_{k'}(t_2-\tau_2)}\, D_{jk'}(\tau_2)\, d\tau_2$$

The correlation function between two stationary random processes x(t) and y(t) is expressed by means of

$$\phi_{xy}(\tau) = \overline{x(t)\, y(t+\tau)} \qquad (11)$$

Thus equation (10) may be rewritten in terms of autocorrelation fuctions and crosscorrelation functions as

$$\phi_{\epsilon_j\epsilon_j}(t_2 - t_1) =$$
$$\sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty W_{jk}(\tau_1)\, d\tau_1 \int_0^\infty [\phi_{s_k s_{k'}}(t_2 - t_1 + \tau_1 - \tau_2) + \phi_{n_k s_{k'}}(t_2 - t_1 + \tau_1 - \tau_2)$$
$$+ \phi_{s_k n_{k'}}(t_2 - t_1 + \tau_1 - \tau_2) + \phi_{n_k n_{k'}}(t_2 - t_1 + \tau_1 - \tau_2)]\, W_{jk'}(\tau_2)\, d\tau_2$$
$$- \sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty D_{jk}(\tau_1)\, d\tau_1 \int_0^\infty [\phi_{s_k s_{k'}}(t_2 - t_1 + \tau_1 - \tau_2) + \phi_{s_k n_{k'}}(t_2 - t_1 + \tau_1 - \tau_2)]$$
$$W_{jk'}(\tau_2)\, d\tau_2 \qquad (12)$$
$$- \sum_{k=1}^{n}\sum_{k'=1}^{n} \int_0^\infty W_{jk}(\tau_1)\, d\tau_1 \int_0^\infty [\phi_{s_k s_{k'}}(t_2 - t_1 + \tau_1 - \tau_2) + \phi_{n_k s_{k'}}(t_2 - t_1 + \tau_1 - \tau_2)]$$
$$D_{jk'}(\tau_2)\, d\tau_2$$

---

* The lower limit in this equation can, of course, be taken as $-\infty$ to include the nonphysically realizable operation of pure prediction, for example, and all subsequent results will follow in like manner.

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(t_2-t_1+\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2 \qquad (12)$$

By setting t_1 equal to t_2, we can determine the mean-squared value of the error. Thus

$$\overline{\epsilon_j^2} = \phi_{\epsilon_j\epsilon_j}(0)$$

$$=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty [\phi_{s_k s_{k'}}(\tau_1-\tau_2)+\phi_{n_k s_{k'}}(\tau_1-\tau_2)$$

$$+\phi_{s_k n_{k'}}(\tau_1-\tau_2)+\phi_{n_k n_{k'}}(\tau_1-\tau_2)]\,W_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty [\phi_{s_k s_{k'}}(\tau_1-\tau_2)+\phi_{s_k n_{k'}}(\tau_1-\tau_2)]\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (13)$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty [\phi_{s_k s_{k'}}(\tau_1-\tau_2)+\phi_{n_k s_{k'}}(\tau_1-\tau_2)]\,D_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

Since $\phi_{xy}(\tau) = \phi_{yx}(-\tau)$, then the above equation can be simplified as

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty [\phi_{s_k s_{k'}}(\tau_1-\tau_2)+\phi_{n_k s_{k'}}(\tau_1-\tau_2)$$

$$+\phi_{s_k n_{k'}}(\tau_1-\tau_2)+\phi_{n_k n_{k'}}(\tau_1-\tau_2)]\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (14)$$

$$-2\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty [\phi_{s_k s_{k'}}(\tau_1-\tau_2)+\phi_{s_k n_{k'}}(\tau_1-\tau_2)]\,W_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

Let us introduce for convenience the notation

$$\phi_{kk'}(\tau)=\phi_{s_k s_{k'}}(\tau)+\phi_{n_k s_{k'}}(\tau)+\phi_{s_k n_{k'}}(\tau)+\phi_{n_k n_{k'}}(\tau) \qquad (15)$$

$$\Theta_{kk'}(\tau)=\phi_{s_k s_{k'}}(\tau)+\phi_{s_k n_{k'}}(\tau) \qquad (16)$$

Then equation (14) can be rewritten in the form

$$\overline{\epsilon^2}=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (17)$$

$$-2\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \Theta_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

This equation forms the starting point for our minimization problem. It gives us a way of measuring the mean-squared error value of the output in terms of the statistical characteristics of the input and the response properties of the actual and ideal systems.

Let us now assume that the inputs to all terminals are uncorrelated with one another. Then all the terms in equation (17) with different subscripts k and k' will drop out. Therefore, equation (17) reduces to

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk}(\tau_1-\tau_2)\,W_{jk}(\tau_2)\,d\tau_2$$

$$-2\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \Theta_{kk}(\tau_1-\tau_2)\,W_{jk}(\tau_2)\,d\tau_2 \qquad (18)$$

$$+\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_k}(\tau_1-\tau_2)\,D_{jk}(\tau_2)\,d\tau_2$$

or

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk}(\tau_2-\tau_1)\,W_{jk}(\tau_2)\,d\tau_2$$

$$-2\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \Theta_{kk}(\tau_2-\tau_1)\,W_{jk}(\tau_2)\,d\tau_2 \qquad (18A)$$

$$+\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_k}(\tau_2-\tau_1)\,D_{jk}(\tau_2)\,d\tau_2$$

where

$$\phi_{kk}(\tau)=\phi_{s_k s_k}(\tau)+\phi_{n_k s_k}(\tau)+\phi_{s_k n_k}(\tau)+\phi_{n_k n_k}(\tau) \qquad (19)$$

$$\Theta_{kk}(\tau)=\phi_{s_k s_k}(\tau)+\phi_{n_k s_k}(\tau) \qquad (20)$$

The next step in the argument is to find a necessary and sufficient condition that $W_{jk}(\tau)$ must satisfy in order that $\overline{\epsilon_j^2}$ will be a minimum. If $W_{jk}(t)$ actually minimizes $\overline{\epsilon_j^2}$ and if we replace $W_{jk}(t)$ by $W_{jk}(t) + a_k A_k(t)$, where a_k is a real number and $A_k(t)$ is an arbitrary function of t, the effect will be to increase $\overline{\epsilon_j^2}$. Thus, for fixed functions $A_k(t)$, $\overline{\epsilon_j^2}$ will be a function of a_k.

$$\overline{\epsilon_j^2}(a_1,a_2,\ldots a_n) = \sum_{k=1}^{n}\int_0^\infty d\tau_1\int_0^\infty d\tau_2\,[W_{jk}(\tau_1)+a_k A_k(\tau_1)][W_{jk}(\tau_2)+a_k A_k(\tau_2)]\,\phi_{kk}(\tau_2-\tau_1)$$

$$-2\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \theta_{kk}(\tau_2-\tau_1)[W_{jk}(\tau_2)+a_k A_k(\tau_2)]\,d\tau_2$$

$$+\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_k}(\tau_2-\tau_1)\,D_{jk}(\tau_2)\,d\tau_2$$

$$=\sum_{k=1}^{n}\int_0^\infty d\tau_1\int_0^\infty d\tau_2\,[W_{jk}(\tau_1)W_{jk}(\tau_2)+a_k A_k(\tau_1)W_{jk}(\tau_2) + a_k A_k(\tau_2)W_{jk}(\tau_1)+a_k^2 A_k(\tau_1)A_k(\tau_2)]\phi_{kk}(\tau_2-\tau_1) \qquad (21)$$

$$-2\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \theta_{kk}(\tau_2-\tau_1)[W_{jk}(\tau_2)+a_k A_k(\tau_2)]\,d\tau_2$$

$$+\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_k}(\tau_2-\tau_1)\,D_{jk}(\tau_2)\,d\tau_2$$

This equation will assume its minimum value when $a_1 = a_2 = \cdots = a_n = 0$. Therefore, the partial derivatives of $\overline{\epsilon_j^2}$ with respect to a_k's must vanish for a_k's = 0, where k = 1, 2, - - - n. If we differentiate equation (21) with respect to a_k's, and then set a_k's = 0, we have

$$\int_0^\infty d\tau_1\int_0^\infty d\tau_2\,[A_k(\tau_1)W_{jk}(\tau_2)+A_k(\tau_2)W_{jk}(\tau_1)]\,\phi_{kk}(\tau_2-\tau_1)$$

$$-2\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \theta_{kk}(\tau_2-\tau_1)A_k(\tau_2)\,d\tau_2 = 0 \qquad (22)$$

$$k = 1, 2, \ldots n$$

Since $\phi_{kk}(\tau_1-\tau_2) = \phi_{kk}(\tau_2-\tau_1)$, equation (22) can be rewritten as

$$2\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk}(\tau_2-\tau_1)\,A_k(\tau_2)\,d\tau_2$$

$$-2\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \theta_{kk}(\tau_2-\tau_1)A_k(\tau_2)\,d\tau_2 = 0$$

or

$$\int_0^\infty A_k(\tau_2)\Big[\int_0^\infty W_{jk}(\tau_1)\phi_{kk}(\tau_2-\tau_1)\,d\tau_1 - \int_0^\infty D_{jk}(\tau_1)\,\theta_{kk}(\tau_2-\tau_1)\,d\tau_1\Big]\,d\tau_2 = 0 \qquad (23)$$

Because this must hold for any function $A_k(t)$, we finally get

$$\int_0^\infty W_{jk}(\tau_1)\phi_{kk}(\tau_2-\tau_1)\,d\tau_1 = \int_0^\infty D_{jk}(\tau_1)\,\theta_{kk}(\tau_2-\tau_1)\,d\tau_1$$

or

$$\int_0^\infty W_{jk}(\tau)\,\phi_{kk}(t-\tau)\,d\tau = \int_0^\infty D_{jk}(\tau)\,\theta_{kk}(t-\tau)\,d\tau \qquad (24)^*$$

$$\text{for } t \geq 0$$

and

$$W_{jk}(\tau) = 0 \qquad \text{for } \tau < 0$$

where k = 1, 2, - - - n.

This is the condition which $W_{jk}(t)$ must satisfy in order that $\overline{\epsilon_j^2}$ will be a minimum under the uncorrelated case. This minimum value is accordingly equal to

$$\overline{(\epsilon_j^2)}_{min} = \sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_k}(\tau_2-\tau_1)D_{jk}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk}(\tau_2-\tau_1)\,W_{jk}(\tau_2)\,d\tau_2 \qquad (25)$$

Solution of Integral Equations Using Complex Variable Techniques for the Uncorrelated Case

Before we start to solve the integral equation (24), there are two important theorems from function theory which we will need in order to get the desired solution.

Theorem I

Let f(t) be an integrable function which vanishes over the time range from t = $-\infty$ to t = 0, and possesses a Fourier transform F(w). Then

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\,e^{-j\omega t}\,dt = \int_0^\infty f(t)\,e^{-j\omega t} \qquad (26)$$

is an analytic and bounded function of the complex variable w in the lower half of the complex plane. Conversely, let F(w) be analytic, bounded, and free from poles in the lower half of the complex plane. Then if f(t) is the inverse Fourier transform of F(w) so that

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\,e^{j\omega t}\,d\omega \qquad (27)$$

---

* If we allow $D_{jk}(\tau)$ to be a physically nonrealizable function, then this equation becomes simply

$$\int_0^\infty W_{jk}(\tau)\,\phi_{kk}(t-\tau)\,d\tau = \int_{-\infty}^{\infty} D_{jk}(\tau)\,\theta_{kk}(t-\tau)\,d\tau$$

It is also true that f(t) vanishes over the time range from t = - ∞ to t = 0. This same argument can be applied to f(t) vanishing over the range (0, ∞) and F(w) being analytic, bounded and free from poles in the upper half of the complex plane.

Theorem II

If G(w) is a positive and real-valued function defined for real values of w for which

$$\int_{-\infty}^{\infty} \frac{|\log|G(\omega)||}{1+\omega^2}\, d\omega < \infty \tag{28}$$

then there exist two functions $G^+(w)$ and $G^-(w)$, such that

$$G(\omega) = G^+(\omega)\, G^-(\omega) \tag{29}$$

is true for all real values of w. Furthermore, $G^+(w)$ is analytic, bounded, and free from zeros and poles in the lower half of the complex plane and $G^-(w)$ is analytic, bounded, and free from zeros and poles in the upper half plane. In other words,

$$G^+(\omega) = \left[G^-(\omega)\right]^* \tag{30}$$

for all real values of w. Here, the asterisk denotes the conjugate quantity.

With these two theorems in mind we can now return to the problem at hand. For convenience equation (24) is rewritten as

$$\int_0^{\infty} W_{jk}(\tau)\,\phi_{kk}(t-\tau)\,d\tau - \int_{-\infty}^{\infty} D_{jk}(\tau)\,\theta_{kk}(t-\tau)\,d\tau = 0 \tag{31}$$

$$\text{for } t \geq 0$$

and

$$W_{jk}(\tau) = 0 \quad \text{for } \tau < 0 \tag{32}$$

where k = 1, 2, - - n.

If equation (31) holds true for all values of t, it can easily be solved by means of Fourier transforms, provided that all the functions $D_{jk}(\tau)$, $\phi_{kk}(\tau)$ and $\theta_{kk}(\tau)$ possess Fourier transforms. However, equation (31) can not be treated in this simple way since the left side is equal to zero only for t equal to or greater than zero. It is evident that in order to fully utilize the Fourier transformation techniques to our final solution, certain modifications of equation (31) must be made before taking the transformation. This modified equation would then be true for all values of t. Finally we transform this modified equation to obtain explicit solution of weighting function.

Since the functions $\phi_{kk}(\tau)$ and $\theta_{kk}(\tau)$, in general, have non-zero values for negative t, the equality of equation (31) is not true for t less than zero. Let us define a function $f_{jk}(t)$ such that

$$f_{jk}(t) = \int_0^{\infty} W_{jk}(\tau)\,\phi_{kk}(t-\tau)\,d\tau - \int_{-\infty}^{\infty} D_{jk}(\tau)\,\theta_{kk}(t-\tau)\,d\tau \tag{33}$$

$$\text{for } t < 0$$

and

$$f_{jk}(t) = 0 \qquad \text{for } t \geq 0 \tag{34}$$

After understanding the nature of this function, $f_{jk}(t)$, we can proceed to combine equations (33) and (34) into one single equation which holds good for all values of t. Thus we have

$$\int_0^{\infty} W_{jk}(\tau)\,\phi_{kk}(t-\tau)\,d\tau - \int_{-\infty}^{\infty} D_{jk}(\tau)\,\theta_{kk}(t-\tau)\,d\tau = f_{jk}(t) \tag{35}$$

$$\text{for all values of } t.$$

Now we can freely take the Fourier transformation on both sides of this equation since the imposed constraint has been removed. The expression on the right side of equation (35) is simply

$$\int_{-\infty}^{\infty} f_k(t)\, e^{-j\omega t}\, dt = \int_{-\infty}^{0} f_k(t)\, e^{-j\omega t}\, dt = \bar{F}_{jk}(\omega) \tag{36}$$

Thus $\bar{F}_{jk}(w)$ can only have poles in the lower half of the complex w-plane. The first term on the left side can be written as

$$\begin{aligned} &\int_{-\infty}^{\infty} e^{-j\omega t}\, dt \int_0^{\infty} W_{jk}(\tau)\,\phi_{kk}(t-\tau)\, d\tau \\ &= \int_{-\infty}^{\infty} \phi_{kk}(t-\tau)\, e^{-j\omega(t-\tau)}\, dt \int_0^{\infty} W_{jk}(\tau)\, e^{-j\omega\tau} \\ &= \pi\, G_{kk}^{\phi}(\omega)\, Y_{jk}(j\omega) \end{aligned} \tag{37}$$

where G_{kk}^{ϕ} is the total power spectral density of input k and $Y_{jk}(jw)$ is the transfer function.

$$G_{kk}^{\phi}(\omega) = G_{s_k s_k}(\omega) + G_{n_k s_k}(\omega) + G_{s_k n_k}(\omega) + G_{n_k n_k}(\omega) \tag{38}$$

The second term on the left side can be treated in the same way.

$$\begin{aligned} &\int_{-\infty}^{\infty} e^{-j\omega t}\, dt \int_{-\infty}^{\infty} D_{jk}(\tau)\,\theta_{kk}(t-\tau)\, d\tau \\ &= \int_{-\infty}^{\infty} \theta_{kk}(t-\tau)\, e^{-j\omega(t-\tau)}\, dt \int_{-\infty}^{\infty} D_{jk}(\tau)\, e^{-j\omega\tau}\, d\tau \end{aligned} \tag{39}$$

$$= \pi G^{\theta}_{kk}(\omega)\,(Yd)_{jk}(j\omega) \qquad (39)$$

where $G^{\theta}_{kk}(\omega) = G_{s_k s_k}(\omega) + G_{n_k s_k}(\omega)$

Thus equation (35) becomes

$$G^{\phi}_{kk}(\omega)\,Y_{jk}(\omega) - G^{\theta}_{kk}(\omega)(Yd)_{jk}(j\omega) = \frac{1}{\pi} F^{-}_{jk}(\omega) \qquad (40)$$

Let us examine the term G^{ϕ}_{kk} expressed in equation (38). It is known that the auto-correlation function of a general stationary random process is an even function and can be expressed or approximated by the expression $\sum_{k=1}^{n} A_k e^{-a_k|\tau|}$

Therefore, auto-power spectrum density is a rational function expressible as the ratio of two polynomials in w^2 with real coefficients. It can be shown that its poles and zeros are placed symmetrically with respect to both real and imaginary axes. Furthermore, we observe that since

$$G_{s_k n_k}(-\omega^*) = \frac{1}{\pi}\int_{-\infty}^{\infty} \phi_{s_k n_k}(\tau)\, e^{j\omega^*\tau}\, d\tau$$

we have

$$G_{s_k n_k}(-\omega^*)^* = \frac{1}{\pi}\int_{-\infty}^{\infty} \phi_{s_k n_k}(\tau)\, e^{-j\omega\tau}\, d\tau$$

$$= G_{s_k n_k}(\omega) \qquad (41A)$$

Thus, the zeros and poles of $G_{s_k n_k}(w)$ are symmetrically placed about the imaginary axis. This same argument can be applied to $G_{n_k s_k}(w)$. In addition, since

$$G_{s_k n_k}(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \phi_{s_k n_k}(\tau)\, e^{-j\omega\tau}\, d\tau$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \phi_{n_k s_k}(-\tau)\, e^{j\omega(-\tau)}\, d\tau$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \phi_{n_k s_k}(\tau)\, e^{j\omega\tau}\, d\tau \qquad (41B)$$

$$= G_{n_k s_k}(\omega^*)^*$$

It follows that zeros and poles of $G_{s_k n_k}(a)$ are the complex conjugates of those of $G_{n_K s_K}(w)$. Therefore, the zeros and poles of the function $G_{s_k n_k}(w) + G_{n_k \bar{s}_k}(w)$ are also located symmetrically about both the real and imaginary axes. Finally, the function $G^{\phi}_{kk}(w)$ has zeros and poles located symmetrically with respect to both real and imaginary axes and is a rational function of w^2. Therefore, $G^{\phi}_{kk}(w)$ will meet the requirement of Theorem II and can be factored into two functions in the form

$$G^{\phi}_{kk}(\omega) = G^{+}_{kk}(\omega)\, G^{-}_{kk}(\omega) \qquad (42)$$

where $G^{+}_{kk}(w)$ and $G^{-}_{kk}(w)$ satisfy the following conditions:

1. The zeros and poles of $G^{+}_{kk}(w)$ lie entirely in the upper half plane and coincide with the zeros and poles of $G^{\phi}_{kk}(w)$ there.

2. $G^{-}_{kk}(w)$ has complementary properties.

Substituting equation (42) into equation (40), we shall have

$$G^{+}_{kk} G^{-}_{kk} Y_{jk} - G^{\theta}_{kk}(Yd)_{jk} = \frac{1}{\pi} F^{-}_{jk}$$

or

$$G^{+}_{kk}\, Y_{jk} = \frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}} + \frac{1}{\pi}\,\frac{F^{-}_{jk}}{G^{-}_{kk}} \qquad (43)$$

Here, for simplicity, the argument w is omitted, and Y_{jk} and $(Yd)_{jk}$ are expressed in terms of w rather than jw.

Since Y_{jk} is a physically realizable function, its poles and zeros are symmetrically placed about the imaginary axis with no poles in the lower half of w-plane. Hence the left side of equation (43) can only have poles in the upper half plane. The first term on the right side can have poles over the entire plane, but the second term only possesses poles in the lower half plane. Thus we have

$$G^{+}_{kk}\, Y_{jk} = \left[\frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}}\right]^{+}$$

or

$$Y_{jk} = \frac{1}{G^{+}_{kk}}\left[\frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}}\right]^{+} \qquad k=1,2,\ldots n \qquad (44)$$

Here the symbol $[\;]^{+}$ means the total upper half plane poles of the function $\frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}}$

If this function is a rational function, the procedure for obtaining its poles in the upper half plane is rather simple. All that we must do is to expand it in partial fractions and throw away all the terms having poles in the lower half plane. If it is not rational, we may use the equation

$$\left[\frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}}\right]^{+} = \int_{0}^{\infty} h(t)\, e^{-j\omega t}\, dt \qquad (45)$$

where

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{G^{\theta}_{kk}(Yd)_{jk}}{G^{-}_{kk}}\right] e^{j\omega t}\, d\omega \qquad (46)$$

This time function h(t) has non-zero values for both positive and negative time. By taking the Fourier transform of h(t) over the positive time interval only, we shall get all its poles in the upper half plane.

Thus, for the uncorrelated case, equation (44) can be used to determine all the system transfer functions directly and independently. With system having n-inputs and m-outputs, there are nxm equations of this type. In frequency domain, the minimum mean-squared error can be expressed as

$$(\overline{e^2})_{min} = \sum_{k=1}^{m}\int_{0}^{\infty} \left|(Yd)_{jk}(\omega)\right|^2 G_{s_k s_k}(\omega)\, d\omega - \sum_{k=1}^{n}\int_{0}^{\infty} \left|Y_{jk}(\omega)\right|^2 G^{\phi}_{kk}(\omega)\, d\omega \qquad (47)$$

General Consideration for the Optimum Multipole Control Systems when Signals and Noises are Correlated.

In this section, the most general case for correlated signals and noises will be considered. They are restricted to stationary processes as in the previous sections. From equation (17), the mean-squared error is given as

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$-2\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \theta_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (48)$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

where $\phi_{kk'}$ and $\theta_{kk'}$ are defined equations (15) and (16) respectively.

Due to the presence of crosscorrelation functions, their associated weighting functions $W_{jk}(\tau_1)$ and $W_{jk'}(\tau_2)$ are tied together, as in the first term of equation (48). Here the two subscripts k and k' can assume different values. Therefore, the situation becomes somewhat more complicated when compared to the uncorrelated case. The usual techniques of the calculus of variation cannot be applied here.

In order to solve this problem, let us define

$$\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,\theta_{kk'}(\tau_1-\tau_2)\,d\tau_1 = \sum_{k=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,\phi_{kk'}(\tau_1-\tau_2)\,d\tau_1 \qquad (49)$$

where $V_{jk}(\tau)$ is an unknown weighting function. Substituting this equation into equation (48) there results

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$-2\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

$$=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (50)$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2 \qquad (50)$$

By adding and subtracting a term

$$\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

the above equation assumes a very useful form

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2 \qquad (51)$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,W_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty W_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

It is very evident that the last four terms can be grouped together to get

$$\overline{\epsilon_j^2} = \sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty\left[V_{jk}(\tau_1)-W_{jk}(\tau_1)\right]d\tau_1\int_0^\infty\left[V_{jk'}(\tau_2)-W_{jk'}(\tau_2)\right]\phi_{kk'}(\tau_1-\tau_2)\,d\tau_2 \qquad (52)$$

$$=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{s_k s_{k'}}(\tau_1-\tau_2)\,D_{jk'}(\tau_2)\,d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)\,d\tau_1\int_0^\infty \phi_{kk'}(\tau_1-\tau_2)\,V_{jk'}(\tau_2)\,d\tau_2$$

$$+\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty\left[V_{jk}(\tau_1)-W_{jk}(\tau_1)\right]d\tau_1\int_0^\infty\left[V_{jk'}(\tau_2)-W_{jk'}(\tau_2)\right]\times$$

$$\overline{\left[S_k(t_1-\tau_1)+n_k(t_1-\tau_1)\right]\left[S_{k'}(t_2-\tau_2)+n_{k'}(t_2-\tau_2)\right]}\,d\tau_2 \qquad (52)$$

The last terms in this expression may be written as

$$\overline{\left\{\sum_{k=1}^{n}\int_0^\infty\left[V_{jk}(\tau_1)-W_{jk}(\tau_1)\right]\left[S_k(t_1-\tau_1)+n_k(t_1-\tau_1)\right]d\tau_1\right\}^2}$$

which shows the fact that this term is always positive. Therefore, to minimize the above equation, we can obviously choose

$$V_{jk}(\tau)=W_{jk}(\tau) \qquad (53)$$

Under this condition, the minimum minimum mean-square error is

$$\overline{\epsilon_{j\min}^2}=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty D_{jk}(\tau_1)d\tau_1\int_0^\infty\phi_{S_kS_{k'}}(\tau_1-\tau_2)D_{jk'}(\tau_2)d\tau_2$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty V_{jk}(\tau_1)d\tau_1\int_0^\infty\phi_{kk'}(\tau_1-\tau_2)V_{jk'}(\tau_2)d\tau_2 \qquad (54)$$

Equation (53) gives not only the necessary condition but also the sufficient one. The reason is that the functions V_{jk}'s defined in equation (49) are expressed in terms of all the known quantities. Therefore, the minimum mean-square error given in equation (54) is not affected by the choice of the system weighting functions. The determination of these optimum weighting functions is thus unique.

By combining equations (49) and (53), we have

$$\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\theta_{kk'}(\tau_1-\tau_2)d\tau_1=\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\phi_{kk'}(\tau_1-\tau_2)d\tau_1 \qquad (55)$$

for $\tau_2\geq 0$

or

$$\sum_{k=1}^{n}\int_0^\infty D_{jk}(\tau_1)\theta_{k'k}(\tau_2-\tau_1)d\tau_1=\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\phi_{k'k}(\tau_2-\tau_1)d\tau_1$$

$$\text{for } \tau_2\geq 0 \qquad (55A)$$

$$k'=1,2,\ldots n$$

and $W_{jk}(t)=0$ for $t<0$. These are the integral equations which $W_{jk}(t)$ must satisfy in order that $\overline{\epsilon_j^2}$ will be a minimum.

In frequency domain, the minimum mean-square error is equal to

$$\left(\overline{\epsilon_j^2}\right)_{min}=\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty(Yd)_{jk}^{*}(\omega)(Yd)_{jk'}(\omega)\,G_{S_kS_{k'}}(\omega)\,d\omega$$

$$-\sum_{k=1}^{n}\sum_{k'=1}^{n}\int_0^\infty Y_{jk}^{*}(\omega)\,Y_{jk'}(\omega)\,G_{kk'}^{\phi}(\omega)\,d\omega \qquad (56)$$

Here the transfer functions Y_{jk}'s we look for must be such that the integration of second summation terms is finite. This condition can be met if each of the fractional functions $Y_{jk}^{*}\,Y_{jk'}\,G_{kk'}^{\phi}$ has denominator with greater power than its numerator.

Solution of Integral Equations for the Correlated Case

The techniques used to solve integral equations for the correlated case are somewhat the same as those for the uncorrelated case. Let us consider the fact that the ideal weighting functions can be physically unrealizable. Equation (55A) can then be rewritten as

$$\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\phi_{k'k}(\tau_2-\tau_1)\,d\tau_1$$

$$-\sum_{k=1}^{n}\int_{-\infty}^{\infty}D_{jk}(\tau_1)\theta_{k'k}(\tau_2-\tau_1)d\tau_1=0 \qquad (57)$$

$$\text{for } \tau_2\geq 0$$

and $W_{jk}(t)=0$ for $t<0$, where $k'=1,2,\ldots n$. The equality of this equation also holds good only for τ_2 equal to or greater than zero. Therefore, a function $f_{jk'}(t)$ can be defined such that

$$f_{jk'}(\tau_2)=\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\phi_{k'k}(\tau_2-\tau_1)\,d\tau_1$$

$$-\sum_{k=1}^{n}\int_{-\infty}^{\infty}D_{jk}(\tau_1)\theta_{k'k}(\tau_2-\tau_1)\,d\tau_1 \qquad (58)$$

$$\text{for } \tau_2<0$$

and

$$f_{jk'}(\tau_2)=0 \qquad \text{for } \tau_2\geq 0 \qquad (59)$$

Combining equations (58) and (59), we shall get a single equation which is true for all values of τ_2. Thus

$$\sum_{k=1}^{n}\int_0^\infty W_{jk}(\tau_1)\phi_{k'k}(\tau_2-\tau_1)\,d\tau_1$$

$$-\sum_{k=1}^{n}\int_{-\infty}^{\infty}D_{jk}(\tau_1)\theta_{k'k}(\tau_2-\tau_1)\,d\tau_1=f_{jk'}(\tau_2) \qquad (60)$$

$$\text{FOR ALL VALUES of } \tau_2$$

Taking the Fourier transformation on both sides of this equation, we obtain

$$\sum_{k=1}^{n}\pi\,G_{k'k}^{\phi}(\omega)\,Y_{jk}(\omega)-\sum_{k=1}^{n}\pi\,G_{k'k}^{\theta}(\omega)(Yd)_{jk}(\omega)=\bar{F}_{jk'}(\omega)$$

or

$$\sum_{k=1}^{n} G^{\phi}_{k'k} Y_{jk} - \sum_{k=1}^{n} G^{\theta}_{k'k} (Yd)_{jk} = \frac{1}{\pi} F^{-}_{jk'} \qquad (61)$$

$$k' = 1, 2, \ldots n$$

Here $F^{-}_{jk'}$ can only possess poles in the lower half plane, and

$$G^{\phi}_{k'k}(\omega) = G_{s_{k'}s_k}(\omega) + G_{s_{k'}n_k}(\omega) + G_{n_{k'}s_k}(\omega) + G_{n_{k'}n_k}(\omega) \qquad (62)$$

$$G^{\theta}_{k'k}(\omega) = G_{s_{k'}s_k}(\omega) + G_{n_{k'}s_k}(\omega) \qquad (63)$$

Thus we have reduced the final solution of our problem to the solution of a set of linear algebraic equations. For each pair of input-output terminals, there are n equations of this type. Hence unique explicit solutions for the n transfer functions associated with a particular terminals-pair can be obtained. m-groups of this type of equations will give the complete solution for the total m x n transfer functions.

Let us write down all the n equations for j terminals-pair as follows:

$$\begin{aligned}
G^{\phi}_{11} Y_{j1} + G^{\phi}_{12} Y_{j2} + \cdots + G^{\phi}_{1n} Y_{jn} &= G^{\theta}_{11}(Yd)_{j1} + \cdots + G^{\theta}_{1n}(Yd)_{jn} + \tfrac{1}{\pi} F^{-}_{j1} \\
G^{\phi}_{21} Y_{j1} + G^{\phi}_{22} Y_{j2} + \cdots + G^{\phi}_{2n} Y_{jn} &= G^{\theta}_{21}(Yd)_{j1} + \cdots + G^{\theta}_{2n}(Yd)_{jn} + \tfrac{1}{\pi} F^{-}_{j2} \\
&\cdots \qquad (64) \\
G^{\phi}_{n1} Y_{j1} + G^{\phi}_{n2} Y_{j2} + \cdots + G^{\phi}_{nn} Y_{jn} &= G^{\theta}_{n1}(Yd)_{j1} + \cdots + G^{\theta}_{nn}(Yd)_{jn} + \tfrac{1}{\pi} F^{-}_{jn}
\end{aligned}$$

In matrix notation,

$$\begin{bmatrix} G^{\phi}_{11} & G^{\phi}_{12} & \cdots & G^{\phi}_{1n} \\ G^{\phi}_{21} & G^{\phi}_{22} & \cdots & G^{\phi}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ G^{\phi}_{n1} & G^{\phi}_{n2} & \cdots & G^{\phi}_{nn} \end{bmatrix} \times \begin{Bmatrix} Y_{j1} \\ Y_{j2} \\ \vdots \\ Y_{jn} \end{Bmatrix} = \begin{Bmatrix} G^{\theta}_{11}(Yd)_{j1} + \cdots + G^{\theta}_{1n}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{j1} \\ G^{\theta}_{21}(Yd)_{j1} + \cdots + G^{\theta}_{2n}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{j2} \\ \cdots \\ G^{\theta}_{n1}(Yd)_{j1} + \cdots + G^{\theta}_{nn}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{jn} \end{Bmatrix} \qquad (65)$$

The desired system transfer functions can then be expressed as

$$\begin{Bmatrix} Y_{j1} \\ Y_{j2} \\ \vdots \\ Y_{jn} \end{Bmatrix} = \frac{1}{G} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} \times \begin{Bmatrix} G^{\theta}_{11}(Yd)_{j1} + \cdots + G^{\theta}_{1n}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{j1} \\ G^{\theta}_{21}(Yd)_{j1} + \cdots + G^{\theta}_{2n}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{j2} \\ \cdots \\ G^{\theta}_{n1}(Yd)_{j1} + \cdots + G^{\theta}_{nn}(Yd)_{jn} + \frac{1}{\pi} F^{-}_{jn} \end{Bmatrix}$$

$$= \frac{1}{G} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} \times \begin{Bmatrix} N^{d}_{j1} + \frac{1}{\pi} F^{-}_{j1} \\ N^{d}_{j2} + \frac{1}{\pi} F^{-}_{j2} \\ \cdots \\ N^{d}_{jn} + \frac{1}{\pi} F^{-}_{jn} \end{Bmatrix} \qquad (66)$$

where

$$G = \begin{vmatrix} G^{\phi}_{11} & G^{\phi}_{12} & \cdots & G^{\phi}_{1n} \\ G^{\phi}_{21} & G^{\phi}_{22} & \cdots & G^{\phi}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ G^{\phi}_{n1} & G^{\phi}_{n2} & \cdots & G^{\phi}_{nn} \end{vmatrix} \qquad (67)$$

$$N^{d}_{jk} = \sum_{k=1}^{n} G^{\theta}_{k'k}(Yd)_{jk} \qquad (68)$$

and $A_{k'k}$ = cofactor of the determinant G
Thus the transfer function for Y_{jk} is given by

$$Y_{jk} = \frac{1}{G} \sum_{k'=1}^{n} A_{k'k} \left(N^{d}_{jk'} + \frac{1}{\pi} F_{jk'} \right) \qquad (69)$$

Let us examine the elements in the determinant G. Along the diagonal with both subscripts being the same, we have

$$G^{\phi}_{kk}(\omega) = G_{s_k s_k}(\omega) + G_{n_k s_k}(\omega) + G_{s_k n_k}(\omega) + G_{n_k n_k}(\omega) \qquad (70)$$

It has been shown in the uncorrelated case that this function G^{ϕ}_{kk} shall be a rational function of w^2 and its poles and zeros are located symmetrically with respect to both real and imaginary axes. The terms $G^{\phi}_{k'k}$ located symmetrically with respect to the diagonal are complex conjugates. Hence equation (67) can be rewritten as

$$G = \begin{vmatrix} G^{\phi}_{11} & G^{\phi}_{12} & \cdots & G^{\phi}_{1n} \\ G^{\phi *}_{12} & G^{\phi}_{22} & \cdots & G^{\phi}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ G^{\phi *}_{1n} & G^{\phi *}_{2n} & \cdots & G^{\phi}_{nn} \end{vmatrix} \qquad (71)$$

Therefore it is a Hermitian determinant. We shall limit ourselves to the case that G does not have zeros on the real axis. It would be true for most of the problems we are interested in. Therefore, G shall be a rational function of w^2 and shall have poles and zeros placed symmetrically with respect to both real and imaginary axes. It is possible then to factor G into two functions in the form

$$G(\omega) = G^{+}(\omega)\, G^{-}(\omega) \qquad (72)$$

where $G^{+}(w)$ and $G^{-}(w)$ satisfy the following

conditions:

1. The zeros and poles of $G^+(w)$ lie entirely in the upper half plane and coincide with the zeros and poles of $G(w)$ there.

2. $G^-(w)$ has complementary properties.

Substituting equation (72) into equation (69) we obtain

$$Y_{jk} = \frac{1}{G^+G^-} \sum_{k'=1}^{n} A_{k'k} \left(N^d_{jk'} + \frac{1}{\pi} F^-_{jk'} \right)$$

or

$$G^+ Y_{jk} = \frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'} + \frac{1}{\pi G^-} \sum_{k'=1}^{n} A_{k'k} F^-_{jk'} \quad (73)$$

Since Y_{jk} is a physically realizable transfer function, it cannot have any pole in the lower half plane. Hence the left side of equation (73) can only possess poles in the upper half plane. The right side of equation (73) can have poles over the entire w-plane. Our problem now is to single out all the poles which are in the upper half plane from the two summation terms.

All the quantities contained in the first summation terms $\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'}$ are known. If this whole function is a rational function, the procedure for obtaining its poles in the upper half plane is simply to expand it in partial fractions and throw away all the terms having poles in the lower half plane. If it is not rational, we have to use the equation

$$\left[\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'} \right]^+ = \int_0^\infty g(t) e^{-j\omega t} dt \quad (74)$$

where

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'} \right] e^{j\omega t} d\omega \quad (75)$$

This time function g(t) has non-zero values for both positive and negative time. By taking the Fourier transform of g(t) over the positive time interval only, we shall get all its poles in the upper half plane.

The second summation terms cannot be handled so easily. However, there are three things we do know. First, the poles of F^-_{jk}' are entirely in the lower half plane and are not known at the present time. Second, the cofactor $A_{k'k}$ shall have poles over the entire plane and these poles are completely known. Third, the zeros and poles of G^- are confined to the lower half plane and they are also completely known. Therefore, it is possible to separate the second summation terms into two groups such that

$$\frac{1}{\pi G^-} \sum_{k'=1}^{n} A_{k'k} F^-_{jk'} = \sum \frac{C_{ie}}{(\omega - \gamma_i)^m} + P^- \quad (76)$$

where

P^- = A function whose poles consist of the poles of F^-_{jk}, the lower-half plane poles of $A_{kk'}$ and the zeros of G^-, and are entirely in the lower half plane.

γ_i = The upper-half plane poles of $A_{k'k}$. These poles are completely known and can appear in multiplicity.

Hence, by combining equations (73), (74) and (76) we shall have

$$G^+ Y_{jk} = \left[\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'} \right]^+ + \sum \frac{C_{ie}}{(\omega - \gamma_i)^m}$$

or

$$Y_{jk} = \frac{1}{G^+} \left\{ \left[\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'} \right]^+ + \sum \frac{C_{ie}}{(\omega - \gamma_i)^m} \right\} \quad (77)$$

$k = 1, 2, \ldots n$

It is evident that, if the inputs between terminals are not correlated, equation (77) can readily be reduced to equation (44).

We have thus expressed Y_{jk} in an explicit form. Everything is known in that equation except the coefficients C_{il}'s which still need to be determined. Here we must be cautious that the transfer functions Y_{jk}'s we solve for will meet the requirement that the second summation terms

$$\sum_{k=1}^{n} \sum_{k'=1}^{n} \int_0^\infty Y^*_{jk}(\omega) Y_{jk'}(\omega) G^\phi_{kk'}(\omega) \, d\omega$$

in equation (55) are finite. The determination of the coefficients C_{il}'s can be achieved by substituting all these transfer functions into any one of the set of equations given in equation (61) to find all the poles in the upper half plane. We shall have

$$\left[\sum_{k=1}^{n} G^\phi_{k'k} Y_{jk} \right]^+ - \left[\sum_{k=1}^{n} G^\theta_{k'k} (Yd)_{jk} \right]^+ = 0 \quad (78)$$

or, more specifically

$$\sum_{k=1}^{n} \int_0^\infty e^{-j\omega t} dt \int_{-\infty}^{\infty} \left[G^\phi_{k'k} Y_{jk} \right] e^{j\omega t} d\omega - \sum_{k=1}^{n} \int_0^\infty e^{-j\omega t} \int_{-\infty}^{\infty} \left[G^\theta_{k'k} (Yd)_{jk} \right] e^{j\omega t} d\omega = 0 \quad (79)$$

From this equation we shall get a set of linear algebraic equations in terms of these coefficients. Therefore, unique values for C_{il}'s can be determined. This procedure will be amplified by the example shown in Appendix A.

In conclusion, the synthesis procedure for determining the optimum transfer function Y_{jk} can be outlined as follows:

1. Factor the function G(w) into two functions $G^+(w)$ and $G^-(w)$.

2. Take out all the upper-half-plane poles from the expression $\frac{1}{G^-} \sum_{k'=1}^{n} A_{k'k} N^d_{jk'}$ and determine all the residues (or coefficients) associated with each pole.

3. Add all the upper-half-plane poles from $A_{k'k}$ with unknown residues to the expression obtained in step #2.

4. Divide the result from step #3 by $G^+(w)$ to get the total expression for $Y_{jk}(w)$.

5. Determine the unknown coefficients by substituting all the transfer functions thus obtained into one of the original system equations and expanding both sides of that equation into partial fractions.

It will be evident that, in general, the poles of the transfer functions consist of two parts. First, all the system transfer functions shall have poles which are the zeros of $G^+(w)$. These poles are completely different from the upper-half-plane

poles in the individual power spectral densities. Secondly, the transfer functions associated with a particular output terminal may contain the upper-half-plane poles from the desired transfer functions $(Yd)_{jk}$ if these poles do nto appear in the power spectral densities. Under very special occasions when part of the zeros and poles in the function $G^{+}(w)$ is cancelled, the system transfer functions shall also contain these cancelled zeros as their poles.

APPENDIX A

Example for the Synthesis of Optimum Multipole Control Systems with Stationary Inputs

As an illustration of the optimalization procedures for multipole control system with stationary inputs, let us consider a system with two inputs and two outputs. The spectral densities are assumed to be

$$G_{s_1 s_1}(\omega) = \frac{1}{\omega^2+1} \qquad G_{s_2 s_2}(\omega) = \frac{2}{\omega^2+2}$$

$$G_{n_1 n_1}(\omega) = 0.2 \qquad G_{n_2 n_2}(\omega) = 0.5$$

$$G_{s_1 s_2}(\omega) = \frac{1}{\omega^2+j\omega+2}$$

and all the other cross-spectral densities are zero. The desired transfer functions of this system are specified as

$$(Yd)_{11} = e^{j\omega} \qquad (Yd)_{12} = 1$$

$$(Yd)_{21} = j\omega \qquad (Yd)_{22} = 1$$

The system configuration is shown in Figure A.

Let us consider the first output terminal. The equations used to solve for Y_{11} and Y_{12} are

$$G_{11}^{\phi} Y_{11} + G_{12}^{\phi} Y_{12} = \left[G_{11}^{\theta}(Yd)_{11} + G_{12}^{\theta}(Yd)_{12}\right] + \frac{1}{\pi} F_{11}^{-}$$

$$G_{21}^{\phi} Y_{11} + G_{22}^{\phi} Y_{12} = \left[G_{21}^{\theta}(Yd)_{11} + G_{22}^{\theta}(Yd)_{12}\right] + \frac{1}{\pi} F_{12}^{-} \quad \text{(A-1)}$$

According to equations (62) and (63), we have

$$G_{11}^{\theta}(\omega) = G_{s_1 s_1}(\omega) = \frac{1}{\omega^2+1}$$

$$G_{12}^{\theta}(\omega) = G_{s_1 s_2}(\omega) = \frac{1}{\omega^2+j\omega+2} = \frac{1}{(\omega+j2)(\omega-j)}$$

$$G_{21}^{\theta}(\omega) = G_{s_2 s_1}(\omega) = G_{s_1 s_2}^{*}(\omega) = \frac{1}{\omega^2-j\omega+2} = \frac{1}{(\omega-j2)(\omega+j)}$$

$$G_{22}^{\theta}(\omega) = G_{s_2 s_2}(\omega) = \frac{2}{\omega^2+2}$$

$$G_{11}^{\phi}(\omega) = G_{s_1 s_1}(\omega) + G_{n_1 n_1}(\omega)$$

$$= \frac{1}{\omega^2+1} + 0.2 = \frac{0.2\omega^2+1.2}{\omega^2+1}$$

$$G_{12}^{\phi}(\omega) = G_{s_1 s_2}(\omega) = \frac{1}{\omega^2+j\omega+2} = \frac{1}{(\omega+j2)(\omega-j)}$$

$$G_{21}^{\phi}(\omega) = G_{s_2 s_1}(\omega) = G_{s_1 s_2}^{*}(\omega) = \frac{1}{\omega^2-j\omega+2} = \frac{1}{(\omega-j2)(\omega+j)}$$

$$G_{22}^{\phi}(\omega) = G_{s_2 s_2}(\omega) + G_{n_2 n_2}(\omega)$$

$$= \frac{2}{\omega^2+2} + 0.5 = \frac{0.5\omega^2+3}{\omega^2+2}$$

From equation (67) the determinant G is given by

$$G = \begin{vmatrix} G_{11}^{\phi} & G_{12}^{\phi} \\ G_{21}^{\phi} & G_{22}^{\phi} \end{vmatrix} = G_{11}^{\phi} G_{22}^{\phi} - G_{12}^{\phi} G_{21}^{\phi}$$

$$= \frac{0.2\omega^2+1.2}{\omega^2+1} \times \frac{0.5\omega^2+3}{\omega^2+2} - \frac{1}{(\omega+j2)(\omega-j)} \times \frac{1}{(\omega-j2)(\omega+j)}$$

$$= \frac{\omega^6+16\omega^4+74\omega^2+124}{10(\omega+j)(\omega-j)(\omega+j1.414)(\omega-j1.414)(\omega+j2)(\omega-j2)}$$

or after determining the roots of the numerator

$$= \frac{(\omega+j3.11)(\omega-j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)}{10(\omega+j)(\omega-j)(\omega+j1.414)(\omega-j1.414)(\omega+j2)(\omega-j2)(\omega-0.481-j1.85)(\omega+0.481-j1.85)} \quad \text{(A-2)}$$

Hence we can factor this G function into two terms:

$$G^{+}(\omega) = \frac{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85)}{10(\omega-j)(\omega-j1.414)(\omega-j2)} \quad \text{(A-3)}$$

$$G^{-}(\omega) = \frac{(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)}{(\omega+j)(\omega+j1.414)(\omega+j2)} \quad \text{(A-4)}$$

(A) Determination of Y_{11}

$$A_{11} = G_{22}^{\phi}(\omega) = \frac{0.5\omega^2+3}{\omega^2+2}$$

$$A_{21} = -G_{12}^{\phi}(\omega) = \frac{-1}{(\omega+j2)(\omega-j)}$$

$$N_{11}^{d} = G_{11}^{\theta}(Yd)_{11} + G_{12}^{\theta}(Yd)_{12}$$

$$= \frac{e^{j\omega}}{\omega^2+1} + \frac{1}{(\omega+j2)(\omega-j)}$$

$$N_{12}^{d} = G_{21}^{\theta}(Yd)_{11} + G_{22}^{\theta}(Yd)_{12}$$

$$= \frac{e^{j\omega}}{(\omega-j2)(\omega+j)} + \frac{2}{\omega^2+2}$$

Thus we have

$$\frac{1}{G^-}\sum_{k'=1}^{2} A_{k'1} N_{1k'}^{d} = \frac{1}{G^-}\left[A_{11}N_{11}^{d} + A_{21}N_{12}^{d}\right]$$

$$= \frac{(\omega+j)(\omega+j1.414)(\omega+j2)}{(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)} \times$$

$$\left\{\frac{0.5\omega^2+3}{\omega^2+2}\left[\frac{e^{j\omega}}{\omega^2+1} + \frac{1}{(\omega+j2)(\omega-j)}\right] - \right.$$

$$\left. \frac{1}{(\omega+j2)(\omega-j)}\left[\frac{e^{j\omega}}{(\omega-j2)(\omega+j)} + \frac{2}{\omega^2+2}\right]\right\} =$$

$$\frac{0.5\omega^4+4\omega^2+10}{(\omega-j)(\omega-j1.414)(\omega-j2)(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)} e^{j\omega}$$

$$+ \frac{0.5(\omega+j)(\omega+j1.414)}{(\omega-j)(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)}$$

$$= (Y_{11})_1 + (Y_{11})_2$$

In order to obtain poles in the upper half plane from this expression we could expand these two terms into partial fractions. Since the first term $(Y_{11})_1$ contains the factor $e^{j\omega}$, equations (74) and (75) must be employed. For the second term $(Y_{11})_2$, we can just discard all the poles in the lower half plane.

$$(Y_{11})_1 = \frac{(0.5\omega^4+4\omega^2+10)}{(\omega-j)(\omega-j1.414)(\omega-j2)(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)} e^{j\omega}$$

$$= \left[\frac{A_1}{\omega-j} + \frac{A_2}{\omega-j1.414} + \frac{A_3}{\omega-j2} + \frac{A_4}{\omega+j3.11} + \frac{A_5}{\omega-0.481+j1.85} + \frac{A_6}{\omega+0.481+j1.85}\right] e^{j\omega}$$

It is evident that the terms which shall be useful in constituting the physically realizable transfer function are the first three terms. Hence we have

$$(g_{11})_1(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{A_1 e^{j\omega}}{\omega-j} + \frac{A_2 e^{j\omega}}{\omega-j1.414} + \frac{A_3 e^{j\omega}}{\omega-j2}\right] e^{j\omega t}\, d\omega$$

$$= A_1 j e^{-(t+1)} + A_2 j e^{-1.414(t+1)} + A_3 j^{3} e^{-2(t+1)}$$

for $t+1 > 0$

so that

$$(Y_{11})_1^{+} = \int_0^{\infty} (g_{11})_1(t)\, e^{-j\omega t}\, dt$$

$$= \frac{A_1 e^{-1}}{\omega-j} + \frac{A_2 e^{-1.414}}{\omega-j1.414} + \frac{A_3 e^{-2}}{\omega-j2} \qquad \text{(A-5)}$$

$$= \frac{0.368 A_1}{\omega-j} + \frac{0.243 A_2}{\omega-j1.414} + \frac{0.135 A_3}{\omega-j2}$$

In a similar way we obtain

$$(Y_{11})_2^{+} = \frac{B_1}{\omega-j} \qquad \text{(A-6)}$$

All the coefficients are found to be:

$$A_1 = -j0.458 \qquad B_1 = -j0.0704$$

$$A_2 = j0.335$$

$$A_3 = -j0.044 \qquad \text{(A-7)}$$

The upper-half-plane poles from A_{11} and A_{21} are $w_1 = j$ and $w_2 = j\,1.414$. Thus we have

$$(Y_{11})_3^{+} = \frac{C_{11}}{\omega-j} + \frac{C_{21}}{\omega-j1.414} \qquad \text{(A-8)}$$

Finally the transfer function Y_{11} is

$$Y_{11} = \left[(Y_{11})_1^{+} + (Y_{11})_2^{+} + (Y_{11})_3^{+}\right] / G^{+}$$

$$= \left[\frac{-j0.458 \times 0.368}{\omega-j} + \frac{j0.335 \times 0.243}{\omega-j1.414} - \frac{j0.044 \times 0.135}{\omega-j2}\right.$$

$$\left. - \frac{j0.0704}{\omega-j} + \frac{C_{11}}{\omega-j} + \frac{C_{21}}{\omega-j1.414}\right] \times$$

$$\frac{10(\omega-j)(\omega-j1.414)(\omega-j2)}{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85)} \qquad \text{(A-9)}$$

$$= \frac{10\left[(C_{11}+C_{21}-j0.3263)\omega^2 - (j3.414C_{11}+j3C_{21}+1.074)\omega - (2.828C_{11} + 2C_{21} - j0.847)\right]}{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85}$$

(B) Determination of Y_{12}

$$A_{12} = -G_{21}^{\phi}(\omega) = \frac{-1}{(\omega-j2)(\omega+j)} \qquad \text{(A-10)}$$

$$A_{22} = G_{11}^{\phi}(\omega) = \frac{0.2\omega^2+1.2}{\omega^2+1} \qquad \text{(A-11)}$$

$$N_{11}^{d} = \frac{e^{j\omega}}{\omega^2+1} + \frac{1}{(\omega+j2)(\omega-j)} \qquad \text{(A-12)}$$

$$N_{12}^{d} = \frac{e^{j\omega}}{(\omega-j2)(\omega+j)} + \frac{2}{\omega^2+2} \qquad \text{(A-13)}$$

Thus we have

$$\frac{1}{G^-}\sum_{k'=1}^{2} A_{k'2}N_{1k'}^{d} = \frac{1}{G^-}\left[A_{12}N_{11}^{d} + A_{22}N_{12}^{d}\right] =$$

$$\frac{(\omega+j)(\omega+j1.414)(\omega+j2)}{(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)} \times$$

$$\left\{\frac{-1}{(\omega-j2)(\omega+j)}\left[\frac{e^{j\omega}}{\omega^2+1} + \frac{1}{(\omega+j2)(\omega-j)}\right] + \frac{0.2\omega^2+1.2}{\omega^2+1}\right.$$

$$\left.\times\left[\frac{e^{j\omega}}{(\omega-j2)(\omega+j)} + \frac{2}{\omega^2+2}\right]\right\} = \qquad \text{(A-14)}$$

$$= \frac{0.2(\omega+j1.414)(\omega+j2)}{(\omega-j2)(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)}\, e^{j\omega}$$

$$+ \frac{0.4\omega^4 + 3\omega^2 + 7.6}{(\omega-j)(\omega-j1.414)(\omega-j2)(\omega+j3.11)(\omega-0.481+j1.85)(\omega+0.481+j1.85)}$$

$$= (Y_{12})_1 + (Y_{12})_2$$

Following the previous procedure, we get

$$(Y_{12})_1^{+} = \frac{0.135A_1'}{\omega-j2} \qquad \text{(A-15)}$$

$$(Y_{12})_2^{+} = \frac{B_1'}{\omega-j} + \frac{B_2'}{\omega-j1.414} + \frac{B_3'}{\omega-j2} \qquad \text{(A-16)}$$

These coefficients are found to be

$$A_1' = -j0.0355 \qquad B_1' = -j0.352 \qquad \text{(A-17)}$$

$$B_2' = j0.268$$

$$B_3' = -j0.0444$$

The upper-half-plane poles from A_{12} and A_{22} are $w_1' = j$ and $w_2' = j2$. Thus we obtain

$$(Y_{12})_3^{+} = \frac{C_{12}}{\omega-j} + \frac{C_{22}}{\omega-j2} \qquad \text{(A-18)}$$

Thus we have

$$Y_{12} = \left[(Y_{12})_1^{+} + (Y_{12})_2^{+} + (Y_{12})_3^{+}\right]/G^{+}$$

$$= \left[\frac{-j0.0355\times0.135}{\omega-j2} - \frac{j0.352}{\omega-j} + \frac{j0.268}{\omega-j1.414} - \frac{0.0444}{\omega-j2}\right.$$

$$\left. + \frac{C_{12}}{\omega-j} + \frac{C_{22}}{\omega-j2}\right] \times$$

$$\frac{10(\omega-j)(\omega-j1.414)(\omega-j2)}{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85)} = \qquad \text{(A-19)}$$

$$10\left[\frac{(C_{12}+C_{22}-j0.1763)\omega^2 - (j3.414C_{12}+j2.414C_{22}+0.6188)}{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85)}\right.$$

$$\left.\omega - (2.828C_{12}+1.414C_{22}-j0.59)\right]$$

(C) Determination of coeffieients

In order to evaluate the coefficients C_{11}, C_{12}, C_{21} and C_{22}, let us substitute equations (A-9) and (A-19) into the first expression of equation (A-1).

$$\frac{0.2\omega^2+1.2}{\omega^2+1} \times \frac{10\left[(C_{11}+C_{21}-j0.3263)\omega^2 - (j3.414C_{11}+j3C_{21}+1.074)\right.}{(\omega-j3.11)(\omega-0.481-j1.85)(\omega+0.481-j1.85)}$$

$$\left.\times\omega - (2.828C_{11}+2C_{21}-j0.847)\right]$$

$$+ \frac{1}{(\omega+j2)(\omega-j)} \times \frac{10\left[(C_{12}+C_{22}-j0.1763)\omega^2 - (j3.414C_{12}+j2.414\right.}{(\omega-j3.11)(\omega-0.481-j1.85)} \qquad \text{(A-20)}$$

$$\left.\times C_{22}+0.6188)\omega - (2.828C_{12}+1.414C_{22}-j0.59)\right]$$

$$(\omega+0.481-j1.85)$$

$$= \left[\frac{e^{j\omega}}{\omega^2+1} + \frac{1}{(\omega+j2)(\omega-j)}\right] + \frac{1}{\pi}F_{11}^{-}$$

Applying equation (79), the partial fractions on both sides of equation (A-20) with the same poles in the upper half plane as denominators must have the same coefficients in the numerators. If there is no such pole on the right side of the equation, the coefficient of that particular term on the left side must be zero. In equation (A-20), there are four poles in the upper half plane; namely,

$$w_1 = j \qquad w_3 = 0.481+j1.85$$

$$w_2 = j3.11 \qquad w_4 = -0.481+j1.85 \qquad \text{(A-21)}$$

It is evident that only the first pole appears on both sides of this equation. Through the evaluation of the residues of these poles, we shall get four independent equations in terms of C_{11}, C_{12}, C_{21} and C_{22}. Thus these constants can be uniquely determined.

The four linear algebraic equations are found to be

$$1.029\,C_{11} + 0.6857\,C_{12} = -j0.272 \tag{A-22}$$

$$0.8757\,C_{11} + 1.09\,C_{21} - 0.9596\,C_{12} - 1.774\,C_{22} = j0.1193 \tag{A-23}$$

$$(1.881 - j0.1364)C_{11} + (0.8402 + j0.0617)C_{21} + (0.7863 - j0.4057)C_{12} + (1.135 + j0.6138)C_{22} = -0.00129 + j0.2293 \tag{A-24}$$

$$(1.881 + j0.1364)C_{11} + (0.8402 - j0.0617)C_{21} + (0.7863 + j0.4057)C_{12} + (1.135 - j0.6138)C_{22} = 0.00129 + j0.2293 \tag{A-25}$$

Combining equations (A-24) and (A-25), there results two simplified independent equations.

$$1.881\,C_{11} + 0.8402\,C_{21} + 0.7863\,C_{12} + 1.135\,C_{22} = j0.2293 \tag{A-26}$$

$$0.1364\,C_{11} - 0.0617\,C_{21} + 0.4057\,C_{12} - 0.6138\,C_{22} = -j0.00129 \tag{A-27}$$

Thus equations (A-22), (A-23), (A-26) and (A-27) can be used to solve for these coefficients.

$$\begin{aligned} C_{21} &= j1.02 \\ C_{12} &= j0.426 \\ C_{22} &= j0.0595 \\ C_{11} &= -j2.548 \end{aligned} \tag{A-28}$$

Finally the transfer function Y_{11} and Y_{12} are

$$Y_{11}(\omega) = 10\,\frac{j0.146\,\omega^2 + 0.115\,\omega + j0.3567}{\omega^3 - j6.81\,\omega^2 - 15.16\,\omega + j11.36}$$

or

$$Y_{11}(p) = -1.46\,\frac{p^2 + 0.788\,p - 2.44}{p^3 + 6.81\,p^2 + 15.16\,p + 11.36} \tag{A-29}$$

and

$$Y_{12}(\omega) = 10\,\frac{j0.299\,\omega^2 + 0.9792\,\omega - j0.699}{\omega^3 - j6.81\,\omega^2 - 15.16\,\omega + j11.36}$$

or

$$Y_{12}(\omega) = -2.99\,\frac{p^2 + 3.27\,p + 2.34}{p^3 + 6.81\,p^2 + 15.16\,p + 11.36} \tag{A-30}$$

The transfer functions Y_{21} and Y_{22} can be evaluated in similar manner.

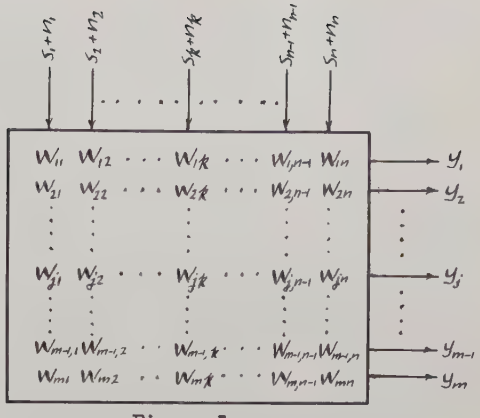


Figure 1

System Diagram

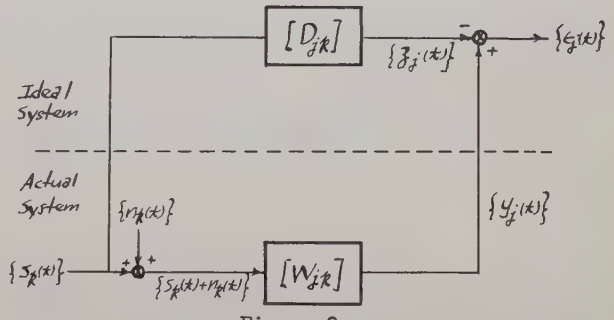


Figure 2

Error Generation Diagram

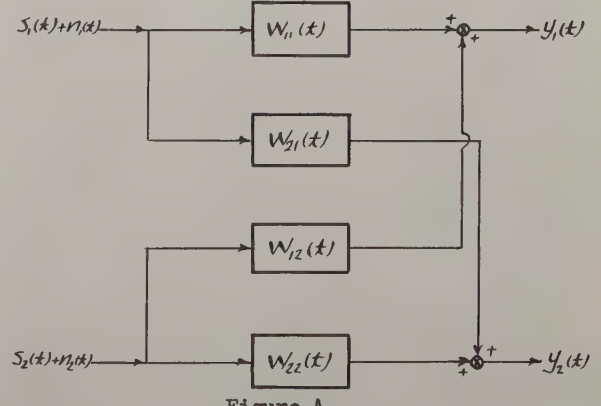


Figure A

Configuration of 2 x 2 Poles System

Reference

1. "A Theory of Multidimensional Servo Systems," M. Golomb and E. Usdin, Journal of the Franklin Institute, pp. 29-51, January 1952.

ON ADAPTIVE CONTROL SYSTEMS*

L. Braun, Jr.

Microwave Research Institute
Polytechnic Institute of Brooklyn
Brooklyn 1, New York

Summary

An attempt is made in the subject paper to evolve a basic philosophy for adaptive control systems.

A method is described for determining the system impulse response from measurements of instantaneous system input and output. The impulse response is expanded in a Taylor series, to facilitate solution of the convolution integral.

From a knowledge of the impulse response and the system error, the necessary correction to the system forcing function is determined, in a manner similar to that used for determination of the impulse response.

The techniques developed are applied to two systems - one stable and one unstable. Curves of the results are presented.

1. Introduction

Ever since the earliest feedback control systems were conceived and built, the designers of such systems have made more and more sophisticated attempts, consciously or not, to imitate the behavior of a human operator when he controls the process under consideration. The comparator, which is common to all feedback systems, is an obvious imitation of the human operator comparing the actual reading of a measuring instrument with the desired reading of that instrument. The output of the comparator, which is the error in the system output, is used to drive the system toward the desired operating condition. This is analogous to the action of the human operator who operates the system control mechanism in a direction and magnitude corresponding to the system error. The anticipating ability of the human being is imitated by the control system designer, when he makes use of a proportional-plus-derivative controller.

A human being is capable of operating an automobile under varying road conditions at a wide range of speeds. This range of variation of operating conditions results in widely differing dynamic behavior of the automobile, and in widely differing control requirements, as the operating conditions change. This places upon the operator the requirement of changing his control action as the operating conditions vary.

It is the purpose of the present paper to develop a technique for designing control systems which imitate this characteristic of the human being. This class of control systems will be called adaptive control systems; where the term "adaptive" has been borrowed from the biological sciences. The biologist calls a system adaptive when its behavior patterns are adjustable depending upon changes in system environment. Many examples of biological adaptive systems are described by Ashby.[1]

A familiar example of a system where a human being applies his adaptability is an automobile. As the velocity of the vehicle and the condition of the road change, the operator must vary the controls in a different manner, in order to achieve desirable performance. The operator senses the characteristics of the automobile from the signals which he applies, and from the various velocity and acceleration components which exist when there are frequent changes of direction, as on a winding road. When an automobile is driven on a long stretch of straight road, no changes of direction, and, consequently, no accelerations occur. The operator then receives insufficient information about the control characteristics of the automobile. Under such circumstances, the normal operator keeps the steering wheel in constant small-amplitude motion, in order to have continuous information about the characteristics of the automobile. Similar small-amplitude motions of the control stick have been observed[2] in studies of pilots of jet aircraft. In both of these cases, the human operator determines the system characteristics from the system forcing functions and system responses, and adapts his control manipulations in such a manner that the vehicle behaves in a desirable manner.

The field of adaptive control systems is very young, and no generally accepted definition of such systems exists. It is therefore necessary to establish a clear definition of adaptive systems, which will be applicable to the control systems to be described in the following pages. The term, adaptive, will be applied to any control system which behaves in a manner similar to the jet-pilot and the automobile operator as described above. Specifically it will be applied to any control system which continuously, or intermittently, measures the impulse response, or some other function which characterizes the system; and which makes use of this system characteristic function to determine, and to generate, the necessary forcing function to cause the system to behave in a desired manner.

*This material is being submitted in partial fulfillment of the requirements for the degree of Doctor of Electrical Engineering at the Polytechnic Institute of Brooklyn.

The next section (2) is devoted to the development of a basic philosophy for adaptive control systems. In subsequent sections, a specific technique is developed for the implementation of this philosophy.

2. A Basic Philosophy for Adaptive Control Systems

The adaptive control system must perform two major tasks:

1. The system impulse response, or some other system characterization, must be determined from measurements of the system forcing function and the system response. This part of the control problem will be called the Identification Problem.

2. The forcing function which is required to obtain a desirable system behavior must be determined from the system characterization. This part of the control problem will be called the Excitation Problem.

The configuration of the systems under consideration here is shown in Fig. 1. The controller includes devices to measure system forcing function and system response; a computer, to determine the characteristic function, and to determine the forcing function required; and signal generating equipment to generate the required forcing function. In addition, the controller will compute the desired system response $c_d(t)$, from the reference input, $r(t)$. The desired output, $c_d(t)$, may in some cases, be the time derivative of $r(t)$; or it may be the response of a reference model which is driven by $r(t)$; or it may be some other desired function of $r(t)$, as required by the specific objectives of the control system under consideration. The computer may be composed of digital, or analog, or a combination of digital and analog components. It does not appear to be necessary to restrict the computer to be digital, as stated in reference 3.

The required forcing function, $m_d(t)$, may be determined from the convolution integral,

$$c_d(t) = \int_{-\infty}^{t} m_d(\tau)\, g(t-\tau)\, d\tau \,. \qquad (1)$$

where,

$c_d(t)$ = desired system output

$m_d(t)$ = forcing function, $m(t)$, which makes $c(t) = c_d(t)$

$g(t)$ = system response to unit impulse applied at $t = 0$.

The integral of Eq. (1) is not valid for non-linear or time-varying systems; since $g(t)$ is assumed, in this equation, to be linear and time-invariant. Adaptive control is not required in systems with linear, time-invariant $g(t)$. In such systems, $g(t)$ may be measured at the outset of the design investigation, and, based upon these measurements, a controller with fixed parameters may be designed using conventional feedback theory. Systems which are non-linear, or which have time-varying parameters, or are both non-linear and time-varying, are the types of systems where adaptive control appears to offer attractive possibilities. The work of succeeding sections is based upon equations of the form of Eq. 1. In order to make this work applicable to time-varying systems, it will be assumed that the system characteristic function and the required forcing function will be re-determined periodically, with a period which is small compared to parameter drift times. Based upon this assumption, parameter variations will be slow enough that the characteristic function may be assumed time-invariant throughout each measuring interval. This work will also be applicable to those non-linear systems whose non-linearities may be represented by piecewise-linear segments, if signal excursions are limited in amplitude.

In order to make use of Eq. 1, to determine $m_d(t)$, the system impulse response must be known. In systems of interest, however, $g(t)$ is not known, a priori, but must be determined from measurements made on the system. The impulse response may be determined from a convolution integral similar to Eq. 1, namely,

$$c(t) = \int_{-\infty}^{t} m(\tau)\, g(t-\tau)\, d\tau \qquad (2)$$

where,

$c(t)$ = actual system response

$m(t)$ = actual system forcing function

Equations 1 and 2 are both integral equations with only a single unknown in each case, $m_d(t)$ and $g(t)$, respectively. In principle, it is a straightforward matter to solve for the unknown; however, because of the folding and scanning implicit in the convolution integral, a solution, even for $c(t)$ in Eq. (2) from a knowledge of $m(t)$ and $g(t)$, is difficult to instrument.[4]

Equations 1 and 2 are easily solved analytically by applying Laplace transform techniques to the equation to transform from the time domain to the complex frequency domain. The instrumentation required to perform this transformation experimentally is cumbersome and slow, so that this approach will not be considered here for application in adaptive control systems. These equations must, consequently, be solved in the time domain. Since exact solution of these equations in the time domain appears to be impossible, a search was made of the literature in the field, to uncover any applicable techniques for obtaining approximate solutions to Eqs. 1 and 2.

Wallman[4] describes a method originally due to Volterra, for approximately solving Eq. 2 by

making measurements of c(t) and m(t) at N discrete instants of time which gives a set of N simultaneous equations for g(t) at these N points. This approach is essentially the one employed by Kalman.[3] Aseltine et al,[5] make use of cross-correlation between the system output and binary noise injected at the input, to obtain the value of g(t) at a number of discrete instants of time. Other approaches to the Identification Problem are described in a survey paper by Aseltine et al.[6]

In the papers by Kalman[3] and by Aseltine et al.,[5] the Excitation Problem is solved by varying parameters in a network in cascade with the plant in such a way that a desirable result is obtained.

Since the integral equation cannot be solved exactly, the Identification Problem must be solved by an approximation technique (Kalman) or by a perturbation scheme (Aseltine). Although there are many cases where the effect of noise injected to permit measurement of g(t) is not objectionable, the approach considered in the following sections employs some approximation technique. Such techniques are applicable to systems where the introduction of noise is intolerable, as well as to systems where perturbation may be used. It is interesting to note that human beings use perturbation only when necessary, as in the operation of an automobile driven on a straight stretch of road. When the normal system inputs provide all the required data, the human operator does not inject noise. In the techniques described in the following sections, it will occasionally be found necessary to inject a disturbance in order to obtain the required data; although during the majority of measurement intervals, the required data will be obtainable with only the desired forcing function exciting the system. The approach used in solving the Excitation Problem will, in each case, be dictated by the approach used in the solution of the Identification Problem.

For the reasons outlined above, the decision was made to use an approximation scheme. The form of the approximation which is used must be appropriate to the system being controlled, to the desired general form of m(t), and to the criterion being used to measure satisfactory performance. Above all, the approximation scheme must be chosen in such a way that the instrumentation of the solutions of the Identification Problem and of the Excitation Problem is easily and economically achievable.

3. Approximation by Maclaurin Series Expansion

3.1 Introduction

Solution of the Identification Problem requires solution of the convolution integral

$$c(t) = \int_{-\infty}^{t} m(\tau)\, g(t-\tau)\, d\tau \qquad (2)$$

for g(t) from known values of c(t) and m(t). If m(t), g(t), and c(t) are each expandable in a Maclaurin series then Eq. 2 may be solved for the coefficients of the Maclaurin series expansion of g(t) in terms of the coefficients of the series for c(t), and the coefficients of the series for m(t). The coefficients of the series for m(t) and c(t) are easily determined from measurements of m(t) and c(t), and their derivatives of various orders.

The infinite lower limit of integration in Eq. 2 indicates a necessity for an infinite memory to store past values of m(t), or at least a memory which has a large enough capacity to store values of m(t) for times which are large compared to the significant system time constants. In order to eliminate the requirement of a large capacity memory, some technique must be developed for eliminating the effect upon c(t) of past excitation.

Since the system parameters are assumed to vary with time, g(t) must be computed periodically, in order to detect any changes which occur. The measurement period, T, must be chosen to be a compromise between two sets of contradictory requirements. The value of T must be chosen small enough that the system parameters may be assumed invariant in any interval of length T. In addition, the smaller the value of T, the fewer terms of the series are required to adequately approximate the functions m(t), c(t) and g(t). On the other hand, as the value of T is made smaller, the measurement problems become more severe. In particular, the measurement of time derivatives becomes more and more difficult in the presence of noise, as the measurement interval becomes smaller and smaller. The extent of the difficulty in choosing a satisfactory value of T depends upon the specifications of the system in each case.

In the remainder of this section techniques are developed for solving the Identification and Excitation Problems without requiring a large capacity memory by using Maclaurin series expansions of m(t), c(t), and g(t). The paper is concluded with two examples illustrating the technique.

3.2 Solution of the Identification Problem Using Maclaurin Series Expansion

In this section, time reference t = 0 will be chosen at the beginning of the control interval of interest. In general, because the series expansions of c(t), m(t), and g(t) must be terminated in a finite (and usually small) number of terms, the applied forcing functions will not be exactly the value needed to force the system output, c(t), to be equal to the desired system output, $c_d(t)$. It is, therefore, necessary to add a correction, $\Delta m(t)$, to the system forcing function at t = 0, in order to force the output to approximate the desired output for $t > 0$. The system forcing function may be written as

$$m(t) = m_1(t) + \Delta m(t), \quad \text{for } t \geq 0 \qquad (3)$$

where,

$m_1(t)$ = system forcing function before correction

$\Delta m(t)$ = correction applied at $t = 0$

For $t<0$, $m(t) = m_1(t)$; i.e., $\Delta m(t) = 0$, for $t < 0$.

Description of $m(t)$, for $t \geq 0$, in terms of the two components, $m_1(t)$ and $\Delta m(t)$, permits solution of the Identification Problem without any requirement for a large capacity memory. If Eq. 3 is inserted in Eq. 2, $c(t)$ becomes

$$c(t) = \int_{-\infty}^{t} m_1(\tau)\, g(t-\tau)d\tau + \int_{o}^{t} \Delta m(\tau)\, g(t-\tau)d\tau = c_1(t) + c_2(t) \tag{4}$$

where

$$c_1(t) = \int_{-\infty}^{t} m_1(\tau)\, g(t-\tau)d\tau \tag{5}$$

$$c_2(t) = \int_{o}^{t} \Delta m(\tau)\, g(t-\tau)d\tau \tag{6}$$

Because $\Delta m(t) = 0$ for $t<0$, solution of Eq. 6 does not require a large capacity memory as do Eqs. 2 and 5. For this reason, attention will be concentrated on Eq. 6. It is important to realize that before Eq. 6 may be utilized, $c_1(t)$ must be removed from $c(t)$, to obtain $c_2(t)$.

The Maclaurin series expansion of $c_2(t)$ for $t>0$ is

$$c_2(t) = C_{20} + C_{21}t + C_{22}\frac{t^2}{2!} + \cdots = \sum_{n=o}^{\infty} C_{2n}\frac{t^n}{n!} \tag{7}$$

where

$$C_{2r} = \left.\frac{d^r c_2(t)}{dt^r}\right|_{t=0_+} = c_2^{(r)}(0_+)$$

From Eq. 4,

$$c_2(t) = c(t) - c_1(t) \tag{8}$$

At the instant $t = 0_+$, Eq. 8 becomes

$$c_2(0_+) = C_{20} = c(0_+) - c_1(0_+) \tag{9}$$

Since $\Delta m(t)$ is the change in $m(t)$ occurring for $t \geq 0$, the forcing function $m_1(t)$ is continuous at $t = 0$, and, therefore, $c_1(t)$ is continuous at $t = 0$, i.e.,

$$c_1(0_-) = c_1(0_+) \tag{10}$$

But

$$c_1(t) = c(t) \text{ ; for } t<0$$

therefore

$$c_1(0_-) = c(0_-) \tag{11}$$

If Eqs. 10 and 11 are inserted in Eq. 9, $c_2(0_+)$ becomes

$$c_2(0_+) = c(0_+) - c(0_-) \tag{12}$$

By similar reasoning,

$$c_2^{(r)}(0_+) = c_{2r} = c^{(r)}(0_+) - c^{(r)}(0_-), \quad \text{for all } r \geq 0 \tag{13}$$

Equation 13 indicates that it is possible to compute the coefficients of the Maclaurin series expansion of $c_2(t)$ by making measurements of the derivatives of $c(t)$ just before and just after the instant $t = 0$; i.e., just before and just after the application of the correction, $\Delta m(t)$. In Eq. 13, the continuity of the stored energy component of the output is employed to eliminate the effect of stored energy on the output, thus eliminating the necessity for a large capacity memory.

The system characterization to be used here in the solution of the Identification Problem is the Maclaurin series expansion of the system impulse response $g(t)$, that is

$$g(t) = G_0 + G_1 t + G_2\frac{t^2}{2!} + \cdots = \sum_{n=o}^{\infty} G_n\frac{t^n}{n!} \tag{14}$$

This expansion is useful since, for small t, only a few terms are required for a good approximation to $g(t)$.

With respect to $m(t)$, it is desirable to choose its form to correspond to that of $g(t)$, in order to simplify the solution of Eq. 6. The form of $m(t)$ to be used here is

$$m(t) = M_0 + M_1 t + M_2\frac{t^2}{2!} + \cdots, \text{ for } t>0 \tag{15}$$

Similarly

$$\Delta m(t) = \Delta M_{-1}\mu_0(t) + \Delta M_0 + \Delta M_1 t + \Delta M_2\frac{t^2}{2!} + \cdots, \tag{16}$$

for $t \geq 0$, where $\mu_0(t)$ = unit impulse applied at $t = 0$

The impulse is added to $\Delta m(t)$ in order to increase the flexibility of the corrective action.

Substitution of Eqs. 14 and 16 in Eq. 6 results in an integrand which is easily integrable term by term. If this integration is carried out, $c_2(t)$ becomes

$$c_2(t) = \Delta M_{-1}G_0 + (\Delta M_{-1}G_1 + \Delta M_0 G_0)t + \\ + (\Delta M_{-1}G_2 + \Delta M_0 G_1 + \Delta M_1 G_0)\frac{t^2}{2!} + \cdots \tag{17}$$

Equations 7 and 17 both express $c_2(t)$ in a power series expansion. In order for these equations to be equal for all t, the coefficients of like powers of t in both series must be equal. This equation of coefficients gives a set of relationships

$$\begin{aligned} C_{20} &= \Delta M_{-1}G_0 \\ C_{21} &= \Delta M_{-1}G_1 + \Delta M_0 G_0 \\ C_{22} &= \Delta M_{-1}G_2 + \Delta M_0 G_1 + \Delta M_1 G_0 \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \end{aligned} \tag{18}$$

Equations 18 may be solved for G_0, G_1, G_2, etc.

$$\begin{aligned} G_0 &= \frac{C_{20}}{\Delta M_1} \\ G_1 &= \frac{C_{21} - \Delta M_0 G_0}{\Delta M_{-1}} \\ G_2 &= \frac{C_{22} - \Delta M_0 G_1 - \Delta M_1 G_0}{\Delta M_{-1}} \end{aligned} \tag{19}$$

and, in general

$$G_r = \frac{C_{2r} - \left[\Delta M_0 G_{r-1} + \Delta M_1 G_{r-2} + \cdots + \Delta M_{r-2} G_1 + \Delta M_{r-1} G_0\right]}{\Delta M_{-1}}$$

for all $r \geq 1$, where $\Delta M_{-1} \neq 0$

If $\Delta M_{-1} = 0$, Eqs. 19 become indeterminate. In fact, in a practical computer, there is a non-zero value of the magnitude of ΔM_{-1} below which the operation of division indicated in Eqs. 19 cannot be carried out. In order to evaluate the coefficients of g(t) under these circumstances, ΔM_{-1} may be made zero in Eqs. 18. Solution of Eqs. 18 for G_0, G_1, etc. in this case yields the relations

$$\begin{aligned} G_0 &= \frac{C_{21}}{\Delta M_0} \\ G_1 &= \frac{C_{22} - \Delta M_1 G_0}{\Delta M_0} \\ G_2 &= \frac{C_{23} - (\Delta M_1 G_1 + \Delta M_2 G_0)}{\Delta M_0} \end{aligned} \tag{20}$$

and, in general,

$$G_r = \frac{c_{2,(r+1)} - \left[\Delta M_1 G_{r-1} + \Delta M_2 G_{r-2} + \cdots + \Delta M_{r-1} G_1 + \Delta M_r G_0\right]}{\Delta M_0}$$

for all $r \geq 1$, where $\Delta M_0 \neq 0$

The correction $\Delta m(t)$ is chosen to make $c(t) = c_d(t)$. It is, therefore, conceivable that a control situation may arise in which the first p coefficients of $\Delta m(t)$ are required to be zero, but the (p+1) coefficient is different from zero. If the results of Eqs. 19 and 20 are generalized to the case where the first p coefficients of $\Delta m(t)$ are zero, the coefficients G_0, G_1, G_2, etc., become

$$\begin{aligned} G_0 &= \frac{C_{2p}}{\Delta M_{p-1}} \\ G_1 &= \frac{C_{2,(p+1)} - \Delta M_p G_0}{\Delta M_{p-1}} \\ G_2 &= \frac{C_{2,(p+2)} - \left[\Delta M_p G_1 + \Delta M_{p+1} G_0\right]}{\Delta M_{p-1}} \end{aligned} \tag{21}$$

and, in general,

$$G_r = \frac{C_{2,(p+r)} - \left[\Delta M_p G_{r-1} + \Delta M_{p+1} G_{r-2} + \cdots + \Delta M_{p+r-2} G_1 + \Delta M_{p+r-1} G_0\right]}{\Delta M_{p-1}}$$

for all $r \geq 1$; where $\Delta M_0 = \Delta M_1 = \cdots = \Delta_{p-2} = 0$, but $\Delta M_{p-1} \neq 0$

The general form of the equation for G_r is the same in Eqs. 21 as it is in Eqs. 19 and 20, so that it is a simple matter to program the computer to use the set of Eqs. 21 which is appropriate to the value of p.

3.3. Solution of the Excitation Problem Using Maclaurin Series Expansion

The form of $\Delta m(t)$ was specified in Eq. 16, in such a way that the Identification Problem was solved relatively easily. The determination of the appropriate values for the coefficients of the series expansion of $\Delta m(t)$ still remains to be considered. These coefficients must be chosen in such a way that the actual output, c(t), closely approximates the desired output $c_d(t)$.

As before, direct use of Eq. 4 requires a large capacity memory, which is undesirable. Using Eq. 4 as a starting point, it is possible to compute the coefficients of the series expansion of $\Delta m(t)$ in a manner similar to that described in Section 3.2 for the computation of the coefficients of the series expansion of $g(t)$.

The output error, $e(t)$, is defined as the difference between $c_d(t)$ and $c(t)$, that is

$$e(t) = c_d(t) - c(t) \tag{22}$$

The correction, $\Delta m(t)$, was defined in Section 3.2 as the correction to be made in $m(t)$ to force $c(t) = c_d(t)$, for $t \geq 0$. Therefore, the required forcing function, $m_d(t)$ is

$$m_d(t) = m_1(t) + \Delta m(t), \quad \text{for } t \geq 0 \tag{23}$$

and inserting Eq. 23 in Eq. 1

$$c_d(t) = \int_{-\infty}^{t} m_1(\tau) g(t-\tau) d\tau + \int_{0^-}^{t} \Delta m(\tau) g(t-\tau) d\tau \tag{24}$$

But, from Eq. 4, the actual output, in the absence of the correction (i.e., $\Delta m(t) = 0$), is

$$c(t) = \int_{-\infty}^{t} m_1(\tau)\, g(t-\tau) d\tau \tag{25}$$

From Eqs. 24 and 25, Eq. 22 becomes

$$e(t) = \int_{0^-}^{t} \Delta m(\ )\, g(t-\ )\, d\ \ \ \

$$\Delta M_{-1} = \frac{E_p}{G_p}$$

$$\Delta M_0 = \frac{E_{p+1} - \Delta M_{-1} G_{p+1}}{G_p} \qquad (30)$$

$$\Delta M_1 = \frac{E_{p+2} - \left[\Delta m_{-1} G_{p+2} + \Delta M_0 G_{p+1}\right]}{G_p}$$

and, in general,

$$\Delta M_r = \frac{E_{p+r+1} - \left[\Delta M_{-1} G_{p+r+1} + \Delta M_0 G_{p+r} + \cdots + \Delta M_{r-2} G_{p+2} + \Delta M_{r-1} G_{p+1}\right]}{G_p}$$

for all $r \geq 1$, where $G_p \neq 0$

The programming of the computer to solve the appropriate set of Eqs. 30, and the generation of the required $\Delta m(t)$ from the solutions of Eqs. 30 are straightforward matters.

3.4 Computer Program for Solution And Excitation Problems

In the illustrative examples of Section 3.5, the series for $c_2(t)$, $g(t)$ and $e(t)$ are each terminated after three terms, so that

$$c_2(t) = C_{20} + C_{21} t + C_{22} \frac{t^2}{2!} \qquad (31)$$

$$g(t) = G_0 + G_1 t + G_2 \frac{t^2}{2!} \qquad (32)$$

$$e(t) = E_0 + E_1 t + E_2 \frac{t^2}{2!} \qquad (33)$$

Three terms are sufficient in each of these series because the value of T was selected (rather arbitrarily here) to be short compared to the significant system time constants. The basis for selection of the value of T is discussed in detail in Section 3.1. The number of terms necessary for good approximations of the functions $c_2(t)$, $g(t)$ and $e(t)$ is determined by the value of T which is selected.

The number of terms in the series for $\Delta m(t)$ is chosen in the examples to be the same as the number of terms in $g(t)$. The correction is composed of an impulse, a step, and a ramp function, so that

$$\Delta m(t) = \Delta M_{-1} \mu_0(t) + \Delta M_0 + \Delta M_1 t \text{ , for } t \geq 0 \qquad (34)$$

In any application of the ideas developed in this chapter, the number of terms used in $\Delta m(t)$ is determined by the number of terms chosen for $g(t)$, by the control action required, and by the economics of the situation at hand.

It is assumed in the illustrative examples that the system impulse response has a first order zero at $s = \infty$, so that $p = 0$ in Eqs. 30. Since $p = 0$, Eqs. 29 may be used directly for computation of the coefficients of $\Delta m(t)$. Coefficients G_0, G_1, and G_2 may be computed from the appropriate set of Eqs. 21.

Calculations of G_0, G_1, G_2, ΔM_{-1}, ΔM_0, and ΔM are performed by the computer every T seconds. In any interval, $nT \leq t \leq (n+1)T$, in which $e(t) = 0$ (i.e., $E_0 = E_1 = E_2 = 0$) there will be no correction. If, however, any one of the coefficients E_0, E_1, or E_2 is different from zero, then a correction is needed. In order to compute the required values of ΔM_{-1}, ΔM_0, and ΔM_1 from Eqs. 29, the values of G_0, G_1, and G_2 must be known. If these are available, then $\Delta m(t)$ is computed without difficulty. Whenever the coefficients of $g(t)$ are not known, $\Delta m(t)$ cannot be determined. Measurement of the coefficients of $g(t)$ requires that the system be in a dynamic condition. In order to enable the computation of G_0, G_1, and G_2 when they are required, but the system is not in a dynamic condition, ΔM_{-1} is made some appropriate value to get the system into a dynamic condition. The values of G_0, G_1, and G_2 obtained in this measurement are then available for the computation of the values of ΔM_{-1}, ΔM_0 and ΔM_1 which are required during the following control interval. This occasional requirement for a perturbation to be introduced is analogous to the occasional requirement for a perturbation of an automobile by its human operator when he receives an insufficient amount of control information, as described in Section 1.

The sequence of calculations carried out by the computer at the instant $t = nT$ is

1. Compute E_0, E_1, and E_2 from Eqs. 22 and 28.

   a. If $E_0 = E_1 = E_2 = 0$, the system behavior is satisfactory, and no correction is necessary.

   b. If any one of the coefficients, E_0, E_1, or E_2 is different from zero, a correction is necessary.

2. If a correction is necessary

   a. and G_0, G_1, and G_2 are known from an earlier measurement, then ΔM_{-1}, ΔM_0, and ΔM_1 are computed from Eqs. 30; and $\Delta m(t)$ is generated and added to $m_1(t)$.

   b. and G_0, G_1, and G_2 are unknown, then ΔM_{-1}, and ΔM_0, and ΔM_1 cannot be computed.

3. If G_0, G_1, and G_2 are unknown and are needed, as in 2b above

   a. they are, easily computed from Eqs. 21 if the controller has applied a correction, $\Delta m(t)$, at $t = nT$.

   b. if there has been no correction applied at $t = nT$, then the controller applies an impulse at $t = nT$, of appropriate amplitude,

$|M_{-1}|$, as a test signal; where $|M_{-1}|$ is chosen sufficiently small so that there will be only an acceptably small disturbance of the system. A disturbance which is "acceptably small" is a disturbance whose amplitude is negligibly small from the viewpoint of desirable system performance. The polarity of this test signal is chosen, each time it is required, to cause the system error to be reduced; that is, if $e(t) > 0$, $M_{-1} > 0$, and if $e(t) < 0$, then $M_{-1} < 0$.

4. Either the test impulse, or the correction, $\Delta m(t)$, applied in step 3 above, permits the calculation of G_0, G_1, and G_2 from Eqs. 21. These values of the coefficients are then available for use in determining the required values of ΔM_{-1}, ΔM_0, and ΔM_1, to be applied to the system during the next control interval, beginning at $t = (n+1)T$.

The sequence of operations, 1-4 above, is repeated at the start of each control interval.

It should be noted that the values of G_0, G_1, and G_2, to be used at the start of any interval for the determination of $\Delta m(t)$, are computed in the immediately preceding interval, so that the data being used are T seconds old. This makes evident the importance of choosing T small compared to parameter drift times.

3.5 Illustrative Examples

The procedure described in Sections 3.2 and 3.3 was applied to a system with a stable impulse response,

$$G(s) = \frac{s + 0.5}{(s+1)(s+2)}$$

and to a system with an unstable impulse response,

$$G(s) = \frac{s + 0.5}{(s+1)(s-1)}$$

In each case, the behavior of the system was investigated by means of a digital computer for the case where $c_d(t)$ was a unit step function. The systems were initially inert. The value of T in both cases was chosen to be 100 milliseconds. This selection was rather arbitrarily made here so that only three terms would be needed in the series for $g(t)$ and $\Delta m(t)$. In a practical application of the techniques presented in this chapter, the value of T must be chosen more carefully, to satisfy the conflicting criteria described in Section 3.1.

The amplitude of the test signal, where it was needed, was chosen to be

$$M_{-1} = \pm 0.01$$

since this amplitude causes output variations which are negligible compared to $c_d(t)$. The (+) sign was chosen when $e(t) > 0$, and the (-) sign was chosen when $e(t) < 0$, in order to tend to reduce the error rather than increase it.

The error signal, $e(t)$, and the forcing function, $m(t)$, are shown for $0 < t < 20T = 2$ seconds; for the stable system in Figs. 2 and 3, respectively; and for the unstable system in Figs. 4 and 5, respectively. It is seen in Fig. 2 that, with the stable system, the error reduces to a value below 1% of $c_d(t)$ at the end of 0.1 seconds (one period of the computer), and then remains below this value; and, in fact, the maximum error in each cycle is less than that in the previous cycle, after the fourth cycle. The system forcing function, $m(t)$, is seen in Fig. 3 to approach, asymptotically, the value $m = 4$; which is the steady state value of $m(t)$ necessary to cause the steady state output to be $c = 1$, as desired. It is seen in Fig. 4, that in the unstable case, the error reduces to a value below 1.2% of $c_d(t)$ at the end of 0.1 seconds (one period of the computer), and then remains below this value; and, in fact, the maximum error in each cycle is less than that in the previous cycle, after the fourth cycle, just as in the case of the stable system. The system forcing function is seen in Fig. 5 to approach, asymptotically, the value, $m = -2$; which is the steady state value of $m(t)$ necessary to cause the steady state output to be $c = 1$, as desired.

5. Conclusions

A measurement technique has been presented as a particular solution to the design of adaptive control systems. The two illustrative examples presented, indicate that good control action is achieved. There are, however, a number of problems of considerable importance in adaptive control systems, which have not, to the author's knowledge been dealt with in the literature in this field. Some of these problems are:

1. What are the effects of computation time and computer errors on system performance?

2. Under what conditions of operation, if any, may the system become unstable?

3. How may the characteristics of the system response (e.g., overshoot, rise time, gain magnitude, phase shift, etc.) to special classes of input functions, such as step inputs or sinusoidal inputs, be rapidly evaluated?

All of the foregoing questions must be answered before the design of adaptive control systems will be established upon as firm a foundation as the design of more conventional linear feedback systems.

6. Acknowledgments

The author wishes to acknowledge his indebtedness to Dr. J.G. Truxal and Dr. E. Mishkin, both of the Polytechnic Institute of Brooklyn, for their active encouragement and many suggestions during the course of this work. Thanks are due to Mr. K. Lian and Mr. A. Zeitlin of the Computing Section of the Microwave Research Institute for programming the computer and obtaining the data used to plot Figs. 2, 3, 4, and 5. This work was

sponsored, in part, by the Office of Ordnance Research, under Contract No. DA-30-069-ORD-1560.

REFERENCES

(1) Design for a Brain, (book), W. Ross Ashby, John Wiley & Sons, Inc., New York, N.Y.

(2) Informative Feedback in Jet-Pilot Control Stick Motion, N.D. Diamantides, AIEE Transactions, Vol. 76, Part II, November 1957, pp. 243-249.

(3) Design of a Self-Optimizing Control System, R.E. Kalman, ASME PaperNo. 57-IRD-12, 1957, Trans. ASME, February 1958.

(4) Electronic Integral Transform Computer and the Practical Solution of Integral Equations, H. Wallman, Journal of the Franklin Institute, Vol. 250, July 1950, p. 45.

(5) A Self-Adjusting System for Optimum Dynamic Performance, J.A. Aseltine, et al., 1958 IRE National Convention Record, Part 4.

(6) A Survey of Adaptive Control Systems, J.A. Aseltine, et al., IRE Transactions on Automatic Control, PGAC-6, December 1958, pp. 102-108

(7) Transients in Linear Systems, Vol. 1, (book), M.F. Gardner and J.L. Barnes; John Wiley & Sons, Inc., New York, N.Y.

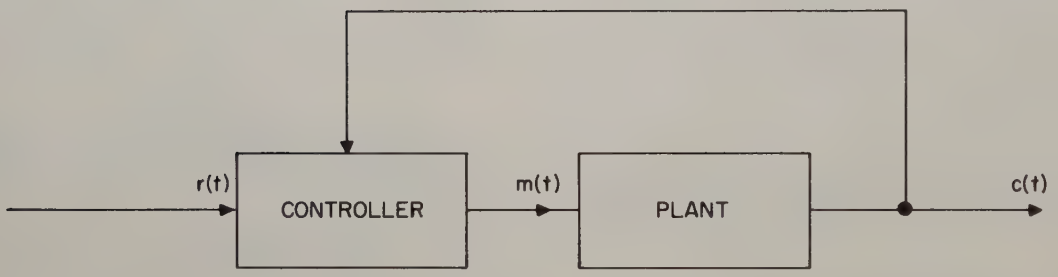
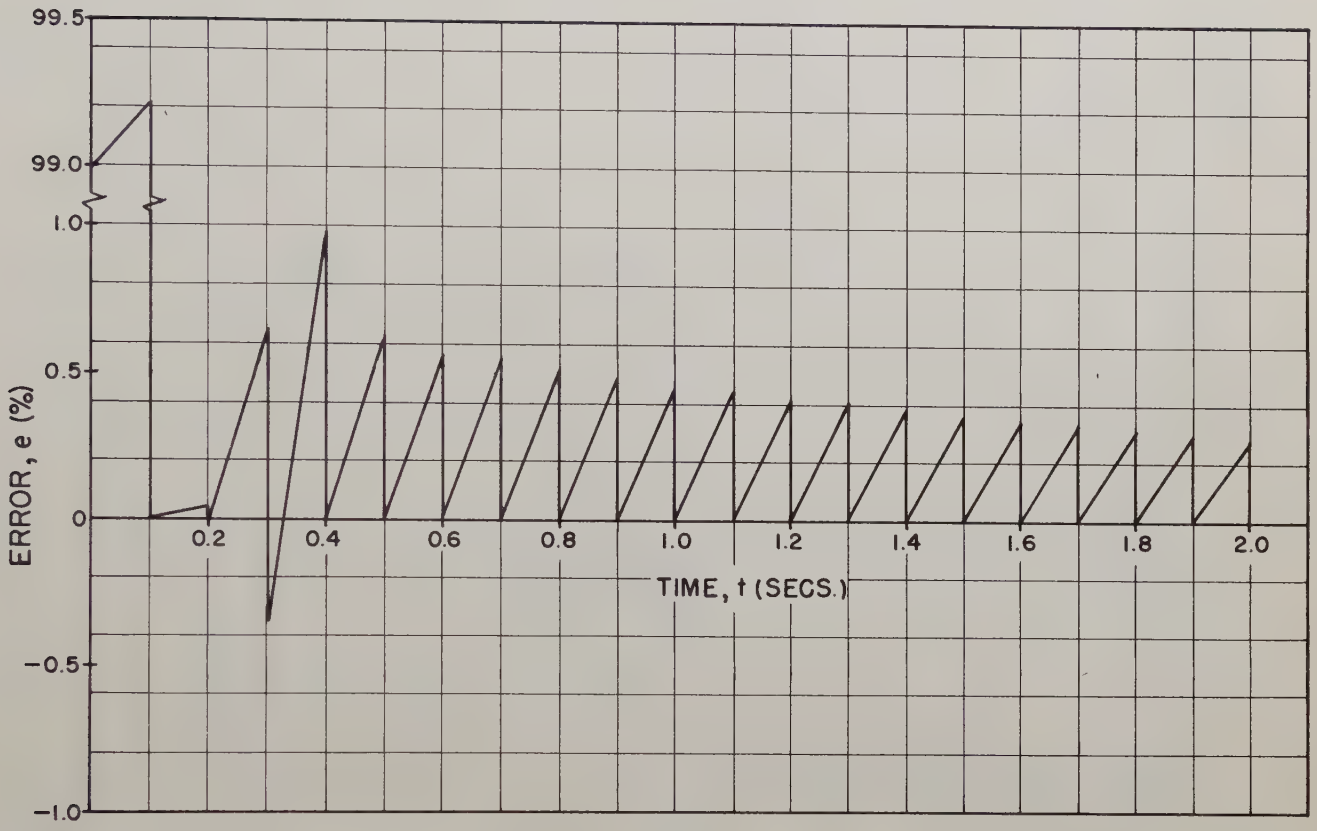


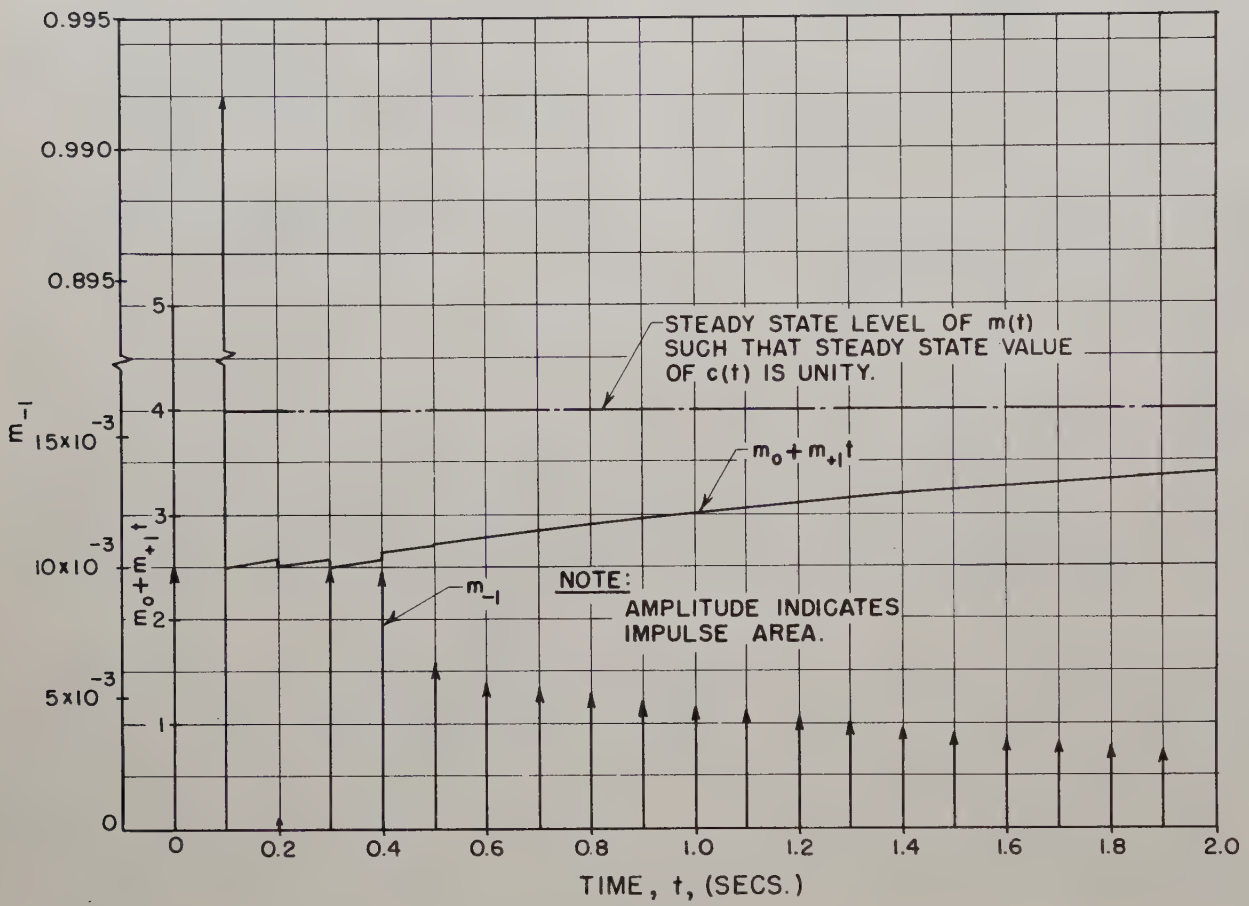
FIG. 1 CONFIGURATION OF ADAPTIVE CONTROL SYSTEM



$$c_d(t) = u_{-1}(t)$$

$$G(S) = \frac{(S+0.5)}{(S+1)(S+2)}$$

FIG. 2 ERROR FUNCTION FOR STABLE SYSTEM



$$c_d(t)=u_{-1}(t)$$

$$G(S)=\frac{(S+0.5)}{(S+1)(S+2)}$$

FIG. 3 FORCING FUNCTION FOR STABLE SYSTEM

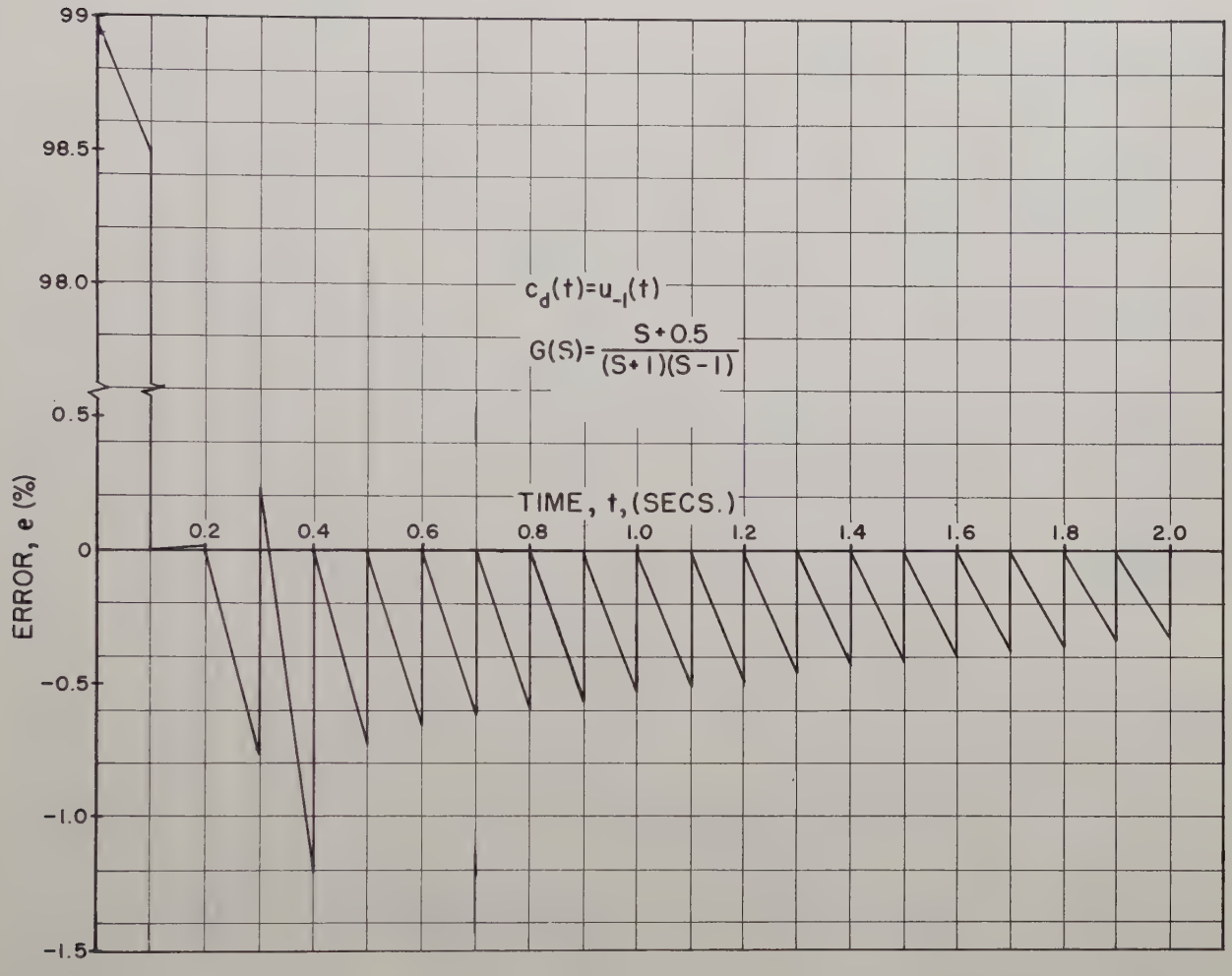


FIG. 4 ERROR FUNCTION FOR UNSTABLE SYSTEM

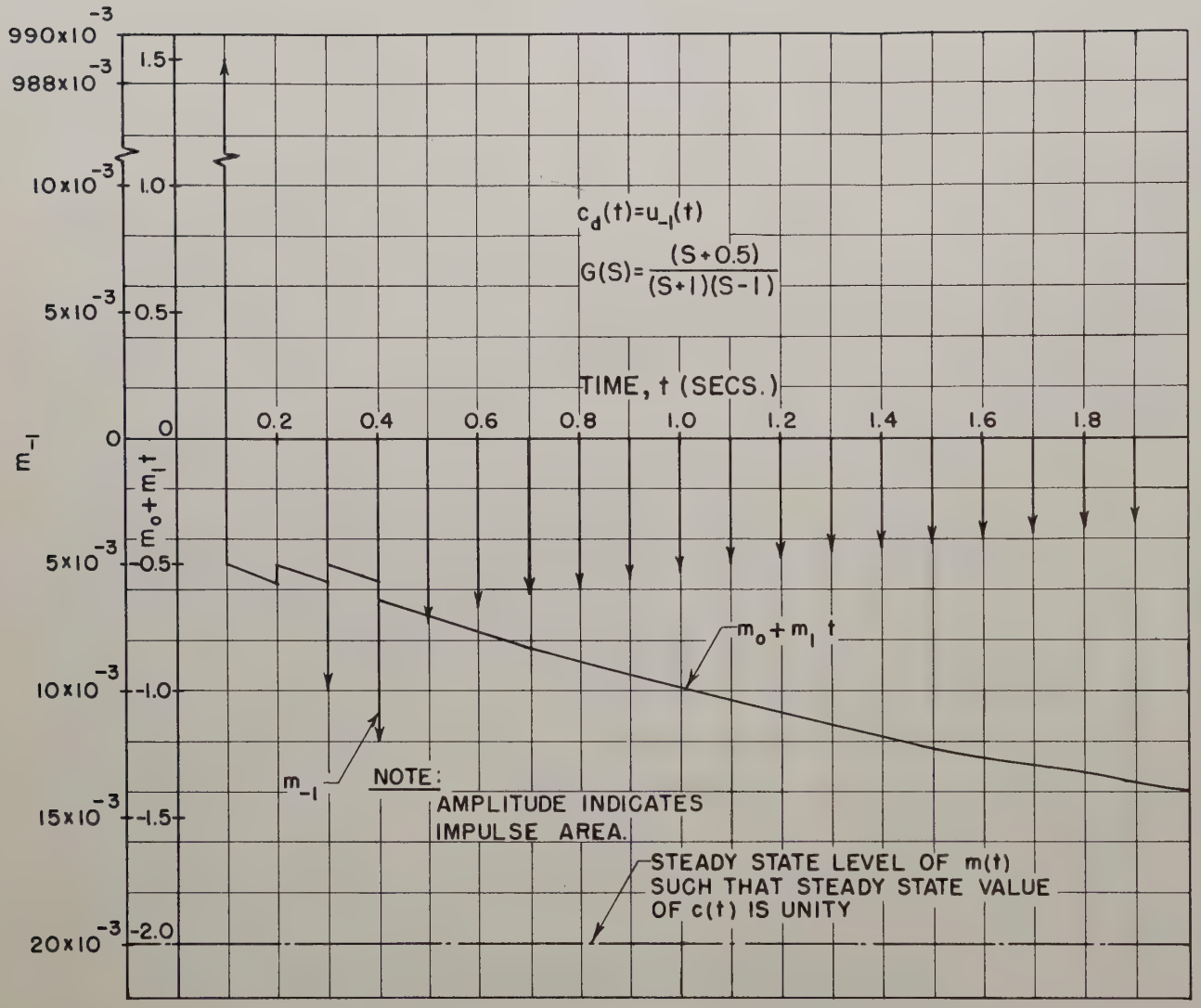


FIG. 5 FORCING FUNCTION FOR UNSTABLE SYSTEM

EXTENSION OF PHASE PLANE ANALYSIS TO QUANTIZED SYSTEMS

Phillip H. Ellis

Countermeasures Division

Sperry Gyroscope Company

Division of Sperry Rand Corporation

Great Neck, L.I., New York

Summary

The increasing applications of numerical control of processes have created a need for new methods of synthesis of control equipment. The method presented is applicable to systems commanded by discretely valued inputs, and processes whose outputs may be similarly quantized. Periodic sampling is not required. The most suitable sampling is by transmission of only significant data, as the new value obtained when the data are changed by a given increment. In certain cases, transmission of data by this means can be used to increase channel capacity. When the data are so quantized, the error signal is constrained to a finite number of discrete values, each of which may be associated with an area in the phase plane. Within each such area, the trajectories of any process subject to phase plane representation are a family of parallel curves. Thus, analytic synthesis may be simplified in the case of certain nonlinear processes. Graphical design is facilitated without requiring deduction of a mathematical representation of the process.

The method is illustrated by synthesis of several systems involving a simple linear process.

INTRODUCTION

The increasing applications of numerical control of processes have created a need for new methods of synthesis of control equipment. To cite a practical example, a tool may be positioned in accordance with data recorded digitally on a paper tape.

In a quantized data channel the exact values of the data are approximated by selecting the nearest of a set of fixed values, as in "rounding off" a number. In transmitting data through such a channel it is not necessary to sample periodically, but only when quantized data change from one possible value to the next. If the signal is not continuously changing, the channel might be made available intermittently for other data without increasing its bandwidth. The particular transmission method assumed in this paper transmits new values of data instantaneously when the least significant digit of the rounded-off data changes. A means of storing the last value transmitted is also provided.

A system using quantized data transmission is illustrated with the aid of Fig. 1, in which G is the process whose control is desired and C is its controller. C is such that m depends on quantized values of r and c; e.g. r may be transmitted from a numerical file such as punched paper tape and c may be transmitted by a numerically indicating measuring device, though functionally such quantizations are included in C. C operates upon the quantized r and c in cuch a way that m is similarly quantized, and any smoothing operations are contained within G. Thus m is constrained to a set of discrete values. G can be represented in the phase plane if it is such that for each value of m, c depends upon no higher than the second derivative of itself with respect to time. If in addition G is an integrating process such that when m is constant, ċ is independent of c, the phase plane consists of a number of regions corresponding to values of m in which the trajectories are parallel.

This application of phase plane analysis is illustrated in the following sections by first demonstrating its use with an on-off control loop, and then by improving the loop's performance by manipulations using the phase plane. Examples of compensator design by both analytical and graphical methods are included. The only generalization involves the applicability of the graphical method. The use of a linear process in these illustrations is for convenience in the introduction and to aid in the transition from analytical to graphical methods.

In all of the illustrations, the system is intended to respond to a step with a final error magnitude less than one-half the incremental unit of quantization, and shall have no oscillation in the steady state. Driving functions other than this step are not considered. The effect after quantization of any other time-varying signals is a succession of superimposed steps. Although these can be handled in the phase plane, it is doubtful that the mathematical labor is justifiable. Therefore, such functions will not be considered.

THE UNCOMPENSATED SYSTEM

A system in the form of Fig. 1, consisting of a continuous linear process and a controller that quantizes its output and

compares it with a quantized input, generating an output that is +1, 0, or -1, depending on whether r - c (after quantizing) is positive, zero, or negative, respectively, can be represented by the block diagram of Fig. 2, which is seen to be an ordinary on-off system with a dead zone. It is to be noted that the definition of the controller equates the unit of c, the quantizing increment, with the width of the dead zone, assuming one-to-one quantizing, which is adjusted in the assignment of a numerical value for K. Thus K is numerically equal to the maximum value of $\dot{c}$.

Case I Unidirectional, zero-error response - The system of Fig. 2 is absolutely stable. This will become apparent in the subsequent development. When a step is applied, c will accelerate; and if the step is sufficiently large, $\dot{c}$ will approach K, being practically equal to K when e enters the dead zone. If K and τ are appropriate $\dot{c}$ will be reduced to zero with e = 0. The condition for this response may be derived from the equation of motion in the dead zone, which is

$$\dot{c} = Ke^{-\frac{t}{\tau}}$$

(e is the base of natural logarithms throughout this paper whenever it appears with an exponent.) The change in c during this deceleration must be half the dead zone:

$$\int_0^\infty Ke^{-\frac{t}{\tau}}\,dt = \frac{1}{2}$$

$$K\tau = \frac{1}{2}$$

is a sufficient condition for the desired response.

The time required for this idealized operation is infinite, since

$$e^{-\frac{t}{\tau}}$$

is nonzero for all finite t. For comparison of methods of reducing the response time, and in deference to the practical fact that it will be brought to rest by friction, the epoch of final value will be arbitrarily defined as the instant when the velocity $\dot{c}$ is reduced to 10 per cent of the maximum value, K. In this case deceleration time is defined:

$$\dot{c} = Ke^{-\frac{T}{\tau}} = \frac{K}{10} \qquad (1)$$

so that

$$T = 2.30\,\tau.$$

Similarly, in time T after application of the step,

$$c = \int_0^T K(1 - e^{-\frac{t}{\tau}})\,dt = K[T - \tau(1 - e^{-\frac{T}{\tau}})]$$

If $T >> \tau$, $c = K(T - \tau)$ and the time required for $\Delta c = n$ $(e > 1/2)$ is $n/K + \tau$.

By means of an approximation similar to Equation 1 it can be shown that this value of T is within 5 per cent of 1/K for a step of 3 or more. Thus the time required for a step $|e_o| \geq 3$ is

$$T = \frac{|e_o| - \frac{1}{2}}{K} + 3.3\tau = \frac{|e_o| + 1.15}{K} = (2|e_o| + 2.3)\tau$$

Phase Plane

The phase portrait is the differential equation in c formed by eliminating t from the equations expressing c and $\dot{c}$ in time functions. This is convenient for a linear process and is more generally applicable than more sophisticated methods. In particular, the method of Kalman using singular points is not applicable to this system which has no singular points in the finite plane.[1]

Assuming initially $\dot{c} = 0$ and m = 1,

$$\dot{c} = K(1 - e^{-\frac{t}{\tau}})$$

$$\Delta c = K\tau\left[\frac{t}{\tau} - (1 - e^{-\frac{t}{\tau}})\right]$$

Eliminating t,

$$\Delta c = -K\tau\left[\ln(1 - \frac{\dot{c}}{K}) - \frac{\dot{c}}{K}\right]$$

In accordance with the assumption of an indefinitely large step, $\dot{c} = K$ when e is diminished to 1/2 and when m becomes zero:

[1] This is true in any region of the plane corresponding to zero loop gain, which is the distinguishing characteristic of on-off systems. In this case two of the singular "points" are at infinity and the third is the position axis (e = 0, e indeterminate).

$$\dot{c} = Ke^{-\frac{t}{\tau}}$$

$$\Delta c = K\tau(1 - e^{-\frac{t}{\tau}})$$

Again eliminating t,

$$\Delta c = \tau(K - \dot{c})$$

Although derived from particular initial conditions, it can be readily shown that these trajectories are completely general with an arbitrary constant of integration. Substituting -e for Δc and noting that the trajectories for +m and -m are negative functions, the general forms are

$$e = e_a + K\tau[\ln(1 + \frac{\dot{e}}{K}) - \frac{\dot{e}}{K}] \qquad (m = +1)$$

(2)

$$e = e_a - K\tau[\ln(1 - \frac{\dot{e}}{K}) + \frac{\dot{e}}{K}] \qquad (m = -1)$$

$$e = e_b - \tau\dot{e} \qquad (m = 0) \quad (3)$$

where e_a and e_b are constants which locate the trajectories along the e axis. Equations 2 and 3 represent families of parallel curves.

Fig. 3 is the phase portrait of the Case I system with $e_o = 4$, drawn from the above equations. It is to be noted that the time required for this operation cannot be calculated from

$$T = \int_{e_o}^{0} \frac{de}{\dot{e}}$$

but that the approximation of Equation 1 can be readily applied. This and subsequent portraits do not show the terminal conditions that arrest the system in finite time.

Modified Phase Plane

The response to successive small steps is sketched in Fig. 4, in which c is the abcissa. This shows that the undershoot error depends on the size of the step, and that for successive equal steps the undershoot diminishes to a limit.

Case II Zero-error Response with One Overshoot - It is immediately suggested by Fig. 3 that if K can be increased by a factor greater than 2 without changing τ, a zero-error response may be obtained by permitting one overshoot, and the time required for a large step can be reduced by a factor approaching the change in K. More generally, it will be seen that the product $K\tau$ is increased, implying no requirement for independence of K and τ. Fig. 5 is a sketch of the phase portrait.

$K\tau$ may be found by simultaneous solutions of the trajectories for their intersections. Assuming, as for Case I, that the step is sufficiently large that $\dot{e}$ approaches -K at point 1, Equation 3 can be written:

$$\frac{1}{2} = a + \tau\dot{e}$$

$$a = \frac{1}{2} - K\tau$$

$$e_{1-2} = \frac{1}{2} - K\tau - \tau\dot{e}.$$

At point 2:

$$-\frac{1}{2} = \frac{1}{2} - K\tau - \tau\dot{e}_2$$

$$e_2 = \frac{1 - K\tau}{\tau} \qquad (4)$$

At point 3:

$$e_{3-4} = -\dot{e}$$

$$-\frac{1}{2} = -\tau\dot{e}_3$$

$$\dot{e}_3 = \frac{1}{2\tau} \qquad (5)$$

Since $e_2 = e_3$, Equation 2 can be expressed:

$$\ln(1 - \frac{1 - K\tau}{K\tau}) + \frac{1 - K\tau}{K\tau} = \ln(1 - \frac{1}{2K\tau}) + \frac{1}{2K\tau} \quad (6)$$

$$\ln \frac{2K\tau - 1}{K\tau} + \frac{1 - K\tau}{K\tau} = \ln \frac{2K\tau - 1}{2K\tau} + \frac{1}{2K\tau}$$

$$\ln 2 = -\frac{2K\tau - 1}{2K\tau}$$

$$K\tau = 1.63$$

The time required for a large step may be obtained by summing the components T_{0-1} through T_{3-4}. From Case I,

$$T_{0-1} = \frac{|e_o| - \frac{1}{2}}{K} + \tau$$

To find T_{1-2} the definite integral of $\dot{e}$ may be set equal to Δe:

$$\int_0^{T_{1-2}} Ke^{-\frac{t}{\tau}}\, dt = 1$$

$$T_{1-2} = \tau \ln \frac{\tau K}{\tau K - 1} = 0.95\,\tau$$

To find T_{2-3}:

$$\int_0^{T_{2-3}} [(K - \dot{e}_2)(1 - e^{-\frac{t}{\tau}}) - \dot{e}_2]\, dt = 0$$

could be solved, but this would require solution of a transcendental equation. It is simpler to solve the $\dot{e}(t)$ equation:

$$(K - \dot{e}_2)(1 - e^{-\frac{T_{2-3}}{\tau}}) + \dot{e}_2 = \dot{e}_3$$

$$T_{2-3} = -\tau \ln \left(\frac{K - \dot{e}_3}{K - \dot{e}_2}\right)$$

Equation 5:

$$\dot{e}_3 = \frac{1}{2\tau}$$

Equation 4:

$$\dot{e}_2 = \frac{1 - \tau K}{\tau}$$

$$= \tau \ln 2 = 0.69\,\tau$$

To find T_{3-4} Equation 1 is used except that in this case $\dot{e}$ is initially $\dot{e}_3$.
Equation 3:

$$\dot{e}_{3-4} = \frac{1}{2\tau} e^{-\frac{T_{3-4}}{\tau}} = \frac{K}{10}$$

$$T_{3-4} = \tau(\ln 5 - \ln K\tau) = 1.12\,\tau$$

The total is

$$T = \frac{|e_o| - \frac{1}{2}}{K} + 3.76\,\tau = \frac{|e_o| + 6.13}{K} = \frac{|e_o| + 6.13}{1.63}\,\tau$$

An attempt to extend this case to an overshoot and an undershoot yielded transcendental equations whose solutions did not fall out readily as did Equation 6. $K\tau$ can be estimated by successive approximations using the graphical method illustrated later.

Comparison of Case I and Case II - In most practical cases, K and τ are not entirely independent. In the case of a mechanical system it is expected that the machine is of an economical size and is efficiently utilized, so that available torque is limited. Then the factor K can be adjusted only by changing the amplification between the machine and the load. If the load is pure inertia, the time constant of the process depends on the amplification factor:

$$\tau = \tau_o \left[1 + \left(\frac{K}{K_o}\right)^2 \frac{J_L}{J_M}\right]$$

where $\tau_o = \tau$ when $K = K_o$

$\frac{K}{K_o}$ = change in amplification factor

J_L = load inertia

J_M = motor inertia

The dependence of τ on K can be calculated in a specific case where J_L/J_M is constant, and where gear inertia is either negligible or may be adequately accounted for by including the input pinion in J_M or conceivably where gear inertia is a known function of K. From this dependence it can be determined whether the product $K\tau$ can be made equal to that required for either system, and then whether the time required for a desired e_o can be less than the time specified for the system. This calculation can be facilitated by estimating K slightly greater than e_o/T, then calculating τ from the above equation, and proceeding with successive approximations if a solution exists.

The only case that can be generalized explicitly for comparison is that in which $J_L = 0$. This case occurs in practice in instrument servos in which $(K/K_o)^2\ J_L << J_M$. Assume that $\tau = 1$. For Case I:

$$T = 2|e_o| + 2.3,$$

and for Case II:

$$T = \frac{|e_o| + 6.13}{1.63} = \frac{|e_o|}{1.63} + 3.76$$

The systems are compared graphically in Fig. 6.

The small step response is not as orderly as for Case I, as may be seen in the more complete phase portrait of Fig. 7. The unit step response may be particularly troublesome in a real system having friction. The modified phase portrait of successive unit steps is shown in Fig. 8, a graphical construction in which the steps are numbered chronologically. These responses appear to be nearly haphazard. It would be good practice to provide a minimum step input of three.

DECELERATION BY REVERSAL

Case III - The system is compensated by a nonlinear delay network incorporating delay D such that when the error enters the dead zone, the process is reversed for a fixed time and then shut down. D is such that the shut-down occurs at $e = 0$, $\dot{e} = 0$ if $|\dot{e}| = K$ at the instant the error enters the dead zone. For steps not large enough for the error rate to approach maximum, the process will reverse, then coast to rest. Input m_1 and output m_2 of the network are related on the same time scale in Fig. 9.

If $|\dot{e}| = K$ at the instant when m_1 becomes zero and m_2 reverses its sense,

$$|\dot{e}| = (2\, e^{-\frac{t}{\tau}} - 1)K$$

D is defined:

$$2\, e^{-\frac{D}{\tau}} - 1 = 0$$

$$D = \tau \ln 2 = 0.69\tau$$

To stop at $e = 0$ the change in e in time D must equal 1/2:

$$K \int_0^{\tau \ln 2} (e^{-\frac{t}{\tau}} - 1)\, dt = \frac{1}{2}$$

$$K\tau = 1.63$$

The large-step response time is:

$$\frac{|e_o| - \frac{1}{2}}{K} + \tau + D = \frac{|e_o| - \frac{1}{2}}{K} + 1.69\tau$$

$$= \frac{|e_o| - 2.26}{K}$$

$$= \frac{|e_o| - 2.26}{1.63}\tau$$

The phase portrait is shown in Fig. 10.

In this system the process does not come to rest after a unit step, but oscillates about $|e| = 1/2$. The limit cycle is not shown on the phase portrait because it depends on a more complete description of D. In this case the process is reversed and coasts back to $|e| > 1/2$, out of the dead zone, and drives back in again. The second excursion of e in and out of the dead zone is completed before the lapse of delay D. If the delayed pulse overrides m_1, an oscillation tracing alternately large and small loops in the phase plane results with an amplitude which does not diminish to zero. If the delayed pulse adds to m_1, an oscillation of a series of similar loops results. If m_1 overrides the delayed pulse, a diminishing oscillation results with the final position at $e \to 1/2$ and the oscillation frequency $\to \infty$.

Incidental nonlinearities, such as friction, might be employed to halt the process before it re-enters the $e > 1/2$ zone. Accounting for nonlinearities in the design of this and the previous systems involved merely the use of the appropriate nonlinear differential equation. Similarly, if a brake could be actuated in the required time, this design could be reduced to a nonlinear Case I or Case II system. If it is feasible to measure $\dot{e}$, the process might be shut down when $\dot{e}$ is nearly zero. It is, of course, possible to stabilize this system by constraining inputs to $\Delta r \geq 2$. This could be done quite easily by quantizing r in twice as many increments as are required for resolution and using only even numbers, which requires only that the least significant bit be zero if r is coded in binomial binary or any other weighted code.

Case IV - Computer Compensated Optimum System - Optimum response of the system will be obtained if the process is reversed at the proper instant to reduce error and velocity to zero simultaneously. The large step response of the Case III system is nearly optimum with a simple "computer."

In this case the reversal time will be computed so that the process will coast to rest in the dead zone. A generalized configuration is shown in Fig. 11.

The ideal system computer would calculate from the initial error e_o the time function:[2]

$$\pm\left[S(0) - 2\,S(t_1) + S(t_2)\right]$$

where t_1 and t_2 are defined by the equations for $\dot{e}_f = 0$, $|\Delta e| = |e_o|$:

$$[(1 - e^{-\frac{t_1}{\tau}}) + 1]\, e^{-\frac{t_2 - t_1}{\tau}} - 1 = 0$$

$$\int_0^{t_1} (1-e^{-\frac{t}{\tau}})dt + \int_0^{t_2-t_1} \left[[(1-e^{-\frac{t_1}{\tau}})+1]e^{-\frac{t_1}{\tau}} - 1\right]dt = \frac{e_o}{K}$$

The practical difficulty of causing the process to have zero error and zero rate simultaneously, and of shutting down the process when this is so, is partially alleviated by bringing the process into the dead zone at a value of $\dot{e}$ sufficiently small that it will coast to rest within the dead zone. The problem is illustrated graphically in Fig. 12. The intersection of the trajectories at e_1 is found by solving Equation 2:

$$e_o + K\tau\,[1n\,(1 + \frac{\dot{e}}{K}) - \frac{\dot{e}}{K}] = e_f + K\tau[1n\,(1 - \frac{\dot{e}}{K}) + \frac{\dot{e}}{K}]$$

$$\frac{\dot{e}}{K} = -\sqrt{1 - e^{-\frac{e_o - e_f}{K\tau}}}$$

The error e_1 at which the process is reversed is:

$$e_1 = e_f - K\tau[1n\,(1+\sqrt{1-e^{-\frac{e_o - e_f}{K\tau}}}) - \sqrt{1-e^{-\frac{e_o - e_f}{K\tau}}}] \qquad (7)$$

A second condition that must be satisfied in order to bring the system to rest is illustrated in Fig. 13.[3] In the equation of the reversal trajectory, e_c is the minimum e_f

$$e = e_f - K\tau[1n\,(1 - \frac{\dot{e}}{K}) + \frac{\dot{e}}{K}] \qquad (8)$$

for which the system will come to rest. Since the equation of the terminal trajectory is Equation 3:

$$e = \frac{1}{2} - \tau\dot{e}$$

the point

$$e = \frac{1}{2},\ \dot{e} = -\frac{1}{\tau}$$

is the intersection, and (Equation 8):

$$e = \frac{1}{2} + K\tau\,[1n\,(1 + \frac{1}{K\tau}) - \frac{1}{K\tau}] \qquad (9)$$

The computer in the block diagram of Fig. 11 calculates e_1 according to Equation 7, compares e with e_1, and inverts e when $e = e_1$. Thus operated upon, e is m_1. It would not be feasible to solve Equation 6 "on line" by analog methods, but discrete, precalculated values of e_1 can be selected by quantizing e_o. The resolution required in quantizing is determined by e_c. That is, the precalculated e_1 must be such that

$$e_c < e_f < \frac{1}{2}$$

in Equation 7, evaluated from exact values of e_o.

In the other systems, it is of no consequence if the comparator inputs are quantized.[4] In this case, however, the error at reversal is critical. Fig. 14 illustrates the effect of the unit of quantization being equal to the width of the dead zone. A further effect not shown is due to e_o having a range of values ± 1/2 about an integer. Excluding this, it is seen that reversing the error at any instant when c changes from one quantized value to the next cannot produce a near optimum response.

[2] Suggested by "Posicast" system of Smith

[3] This requirement is less stringent than to bring to rest with e = 0. The convenience of this freedom will be apparent subsequently.

[4] If r and c are constrained to integers, the nonlinearity introduced in Fig. 1 is intrinsic in the comparator.

If the system could be designed to have $e_c < 0$, the desired response could always be obtained for $|e_o| \to \infty$ provided that c were quantized in increments of 1/2. This may be visualized graphically by noting that a given change in e_1 produces a smaller change in e_f for all $|e_o| < \infty$. For decreasing $|e_o|$ the restraint on e_c is more severe, or conversely, for a given e_c smaller quantizing increments permit smaller $|e_o|$. However, if $K\tau \leq 1$, the system reduces to Case I, and the minimum $|e_c|$ from Equation 9 is

$$|e_c|_{min} = \ln 2 - \frac{1}{2} = 0.193$$

DESIGN IN THE PHASE PLANE

The examples illustrated previously have demonstrated the utility of the phase plane as an aid to visualization and its use in facilitating an analytical design in certain cases where other methods are unwieldy or inapplicable. However, two disadvantages became apparent in the course of developing those examples. In an attempt to generalize Case II transcendental equations were encountered. Also, little light has been shed on manipulation of nonlinear processes. Both difficulties may be removed satisfactorily by effecting design by graphical construction in the phase plane. It is necessary only that G be constrained and the regions be defined as stated earlier, which is to say that in each region the dynamic behavior of G is independent of initial conditions. Additional remarks concerning the characteristics of G are given later. An illustration of the graphical method will be given now using the linear process treated in the previous cases.

Equations 2 and 3 may be written as functions of the same variable, $\dot{e}/K$.

Equation 2:

$$\frac{e}{K} = \frac{e_a}{K} + \tau\left[\ln\left(1 + \frac{\dot{e}}{K}\right) - \frac{\dot{e}}{K}\right]$$

$$\frac{e}{K} = \frac{e_a}{K} - \tau\left[\ln\left(1 - \frac{\dot{e}}{K}\right) + \frac{\dot{e}}{K}\right]$$

Equation 3:

$$\frac{e}{K} = \frac{e_b}{K} - \tau\frac{\dot{e}}{K}.$$

Since e_a and e_b are arbitrary, the implicit equations

Equation 2:

$$\frac{e}{K} = \tau\left[\frac{\dot{e}}{K} + f\left(\frac{\dot{e}}{K}\right)\right] \qquad (10)$$

and Equation 3:

$$\frac{e}{K} = -\tau\frac{\dot{e}}{K} \qquad (11)$$

describe the dynamics in the phase plane. Since the phase plane trajectories, Equation 2, are imaged in the point $e = e_a$, $\dot{e} = 0$, it is possible to construct a single pair of curves defining the system dynamics, Equations 10 and 11, where τ and K are scale factors applied to the plane. These curves can be cut in a plastic template to an appropriate scale and design effected by trial layout in the phase plane. Many of the figures in this paper were so drawn.

Let it be assumed that the process is such that no adjustment of K yields the product τK small enough for Cases I, II or III, and the maximum step is specified. To estimate whether a case IV design is possible, adaptation of the equations used to obtain the response time for Case II is made, giving as an approximation to response time:

$$T = \frac{|e_o| - \frac{1}{2}}{K} + 3\tau$$

where T is maximum time of response to maximum step e_o and K and τ are compatible parameters of the process. If a K and τ that satisfy the inequality exist, the design is possible. This is not an expression for optimizing the system for this step, since it assumes that the maximum $\dot{e}$ is obtained, and does not preclude improvement by increasing K in all cases. It is an approximation because the term in includes deceleration by reversal from maximum $\dot{e}$ to rest, and also because the coefficient of τ is rounded to a whole number. However, an initial approximation to the maximum K (and τ) that can be used, as well as the minimum K, can be made.

An attempt will now be made to design a system for which the approximation $\tau K = 6$ is adequate and realizable. Fig. 15 illustrates the procedure used. In this modified phase plane, the initial value of c can have any value in the 0 range. If the final value of c in response to a step of 21 is to be constant in the 21 range, the deceleration must intersect the transition between 20 and 21 with $\dot{e}$ between 0 and $1/\tau$.

If e_o is quantized in unit increments, switching from the accelerating mode to the decelerating mode must occur in the area bounded by the two accelerating curves. As can be seen, for steps of 21 or greater, it is necessary to calculate e_1 and measure e to a greater accuracy than unit quantization. If e is quantized, it must be in units of 1/4, assuming binary quantization.

Working to successively smaller steps, it is seen that for $e_o = 16$, e must be quantized in smaller increments; in this case, 1/16. As e_o is reduced, it is apparent that the increments must be still smaller because the range of e in which switching must occur becomes smaller until at $e_o = 8$, it vanishes. This difficulty may be relieved by quantizing e_o in smaller increments. Since it has been determined that e must be quantized in increments of 1/16, e_o may be also. The smallest switching range for a step of 2 (from the 0 range to the 2 range) is for a step from the highest (leftmost) increment in the 0 range. This switching range is greater than 1/16, so that any step of 2 or greater may be made. However, all steps of 1 cannot be made without still smaller quantizing increments.

Aside from these mathematical difficulties, other highly probable effects serve to render this system impractical. It is acutely sensitive to changes in τ and K, and practical components are not as well behaved as the idealized curves due to bearing irregularities, slot effect, commutation, eccentricities, imperfect gearing and temperature, to name only those pertinent to electric motors. It may be postulated, however, that small transportation lags, if predictable, due to data transmission, comparison, computation, etc., may be compensated in the computation.

APPLICABILITY OF THE METHOD

This extension of phase plane analysis requires that the process operate in defined regions wherein the phase plane loci are parallel, which implies that within a region there is no feedback around the process. It is seen in the examples given that, due to the nature of the nonlinearity, no expression that is valid within a region can be written for the transmission from e to c.[5] There must be effectively a nonlinearity providing a finite number of constant inputs to the process.

[5] A describing function for the nonlinearity cannot be written without a signal of sufficient amplitude to cause a switching action, which accomplishes a transfer to another region.

The process itself must be such that for each input, its behavior is described independent of previous history - reiterating, c depends upon no higher than the second derivative of itself. Or, stated directly and mechanistically, acceleration may be any single-valued function of velocity.

These conditions permit analysis without recourse to approximations such as the straight-line representation of smooth nonlinearities commonly used to obtain describing functions. The phase plane loci of an efficient induction motor may be obtained, for instance, from speed-torque curves and the total inertia; and if additional nonlinearities such as coulomb friction exist, or if the load inertia is a function of velocity (a centrifuge, for instance), the loci may be obtained directly from records of velocity versus time. The latter recourse may be used even in the case of machines such as synchronous electric motors that hunt.

The student of nonlinear control systems will recognize a similarity between this extension and a method of analysis for systems having backlash. The present extension was suggested by that method, and the similarity may suggest other applications not foreseen by the author.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the assistance and advice he received from Professors Eliezer Mishkin and Robert Staffin of the Polytechnic Institute of Brooklyn during the preparation of this paper which was submitted in partial fulfillment of the requirements for the degree of master of electrical engineering.

BIBLIOGRAPHY

1. R.E. Kalman, "Phase Plane Analysis of Automatic Control Systems With Known Linear Gain Elements," Application and Industry, p. 383 (Jan. 1954).

2. O.J.M. Smith, "Posicast Control of Damped Oscillatory Systems," Proc. I.R.E., Vol. 45, No. 9, p. 1249 (Sept. 1957).

3. J.G. Truxal, Control System Synthesis, p. 660, McGraw-Hill Book Publishing Co., Inc., (1955).

4. C.L. Smith and C.T. Leondes, "An Analysis of the Effects of Certain Nonlinearities on Servomechanism Performance," I.R.E. WESCON Convention Record, Part 4, (1957).

5. I. Flugge-Lotz and H.E. Lindberg, "On the Design and Comparison of Contactor Control Systems," I.R.E. WESCON Convention Record, Part 4, 1957).

6. R.K. Richards, Digital Computer Components and Circuits, D. Van Nostrand Co., Inc., (1958).

7. T.W. Tucker, "A System Study for a Digital-to-Analog Servomechanism," MIT Servomechanisms Laboratory, (1955).

8. W.L. Poland, "Investigation of a Servo-Type Pulse-to-Analog Data Converter," MIT Servomechanisms Laboratory, (1951).

9. "Design, Development and Evaluation of a Numerically Controlled Milling Machine," MIT Servomechanisms Laboratory, (1956).

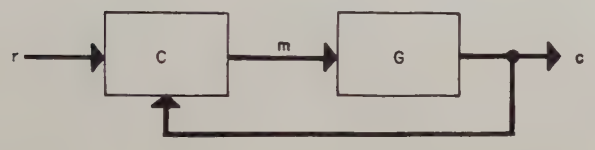


FIG. 1. GENERALIZED SYSTEM

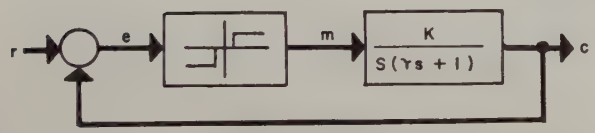


FIG. 2. THE UNCOMPENSATED SYSTEM

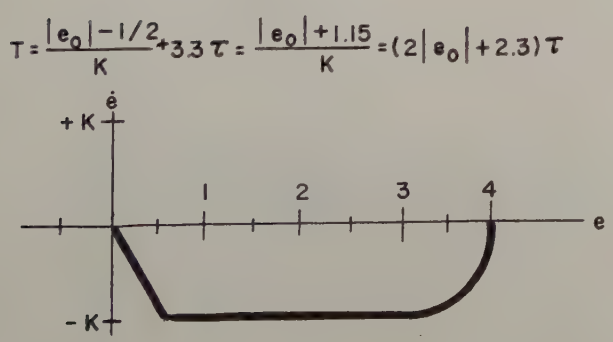


FIG. 3. CASE I SYSTEM FOR $e_0 = 4$

BIOGRAPHY

Phillip H. Ellis, Engineer

Countermeasures Division

Sperry Gyroscope Company
Division of Sperry Rand Corporation
Great Neck, L.I., New York

Phillip H. Ellis, an engineer in the Countermeasures Division, joined the Sperry Gyroscope Company in 1952. He was initially employed as a field engineer for an automatic weapon system, and was assigned to his present position in 1955. He is currently responsible for research engineering for automatic checkout of weapon systems.

Ellis was granted the B.S.E.E. degree by the University of California in 1952. He is currently a candidate for the M.E.E. at the Polytechnic Institute of Brooklyn. He is a senior member of the I.R.E., and is a member of Eta Kappa Nu.

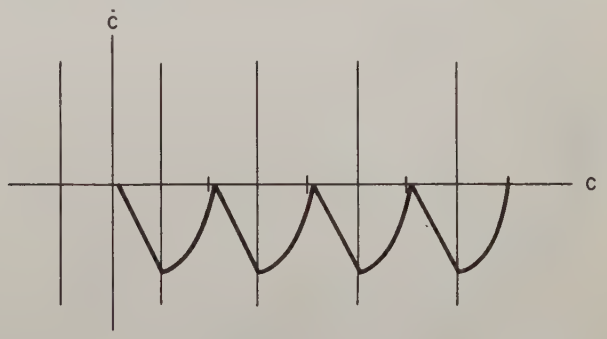


FIG. 4 MODIFIED PHASE PORTRAIT OF CASE I SYSTEM SUCCESSIVE UNIT STEPS

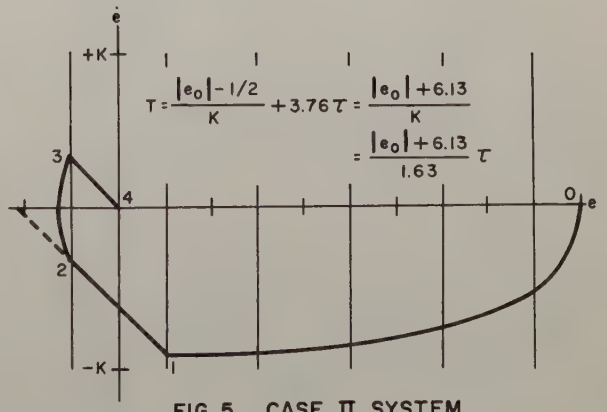


FIG. 5. CASE II SYSTEM

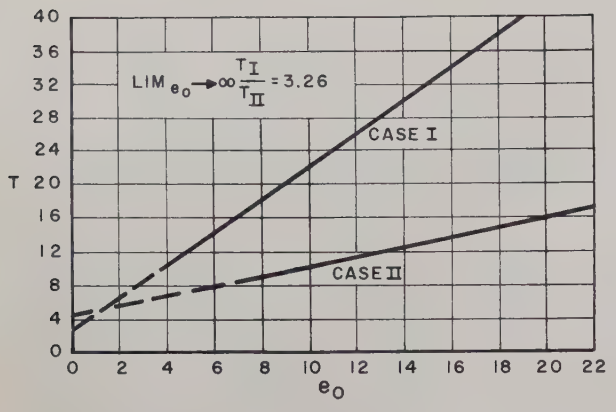


FIG. 6. COMPARISON OF RESPONSE TIME

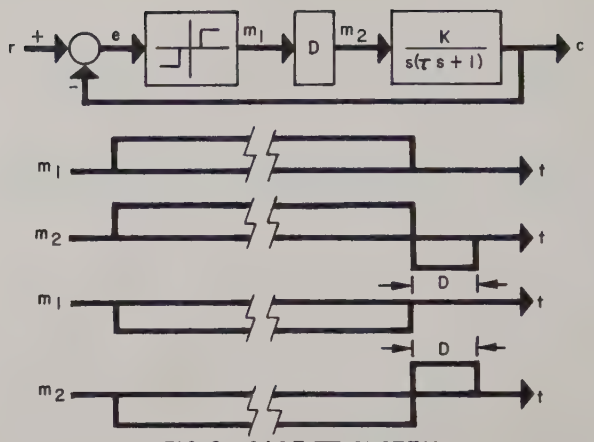


FIG. 9. CASE III SYSTEM

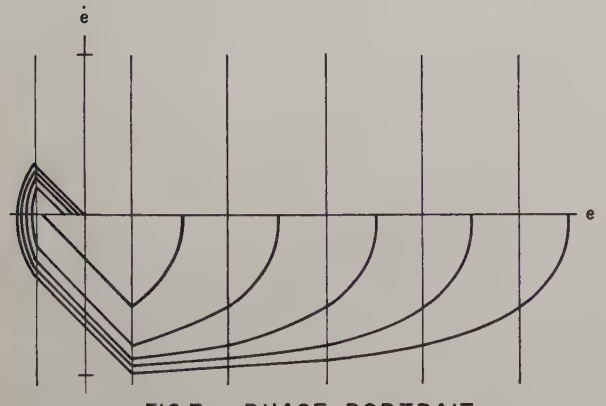


FIG. 7. PHASE PORTRAIT OF CASE II SYSTEM

$$T = \frac{|e_0| - 1/2}{K} + \tau + D = \frac{|e_0| - 1/2}{K} + 1.69\tau = \frac{|e_0| - 2.26}{K} = \frac{|e_0| - 2.26}{1.63}\tau$$

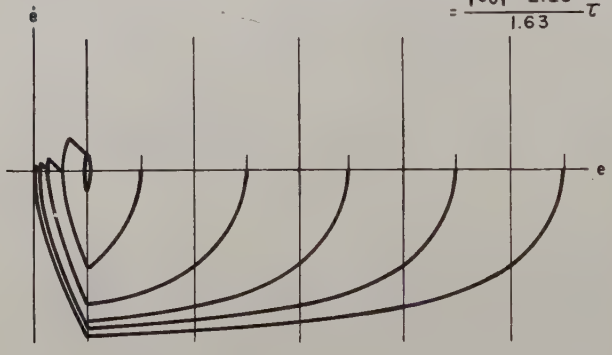


FIG. 10. PHASE PORTRAIT OF CASE III SYSTEM

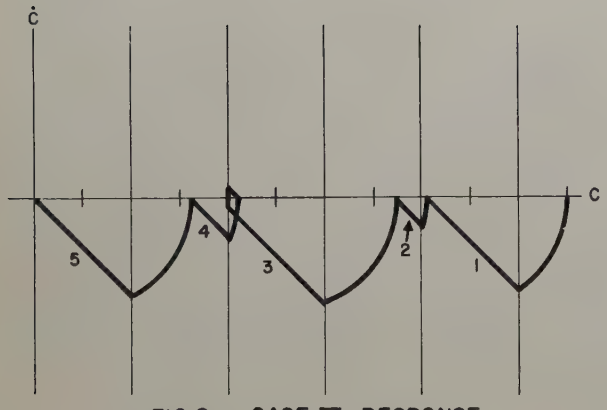


FIG. 8. CASE II RESPONSE TO CONSECUTIVE UNIT STEPS

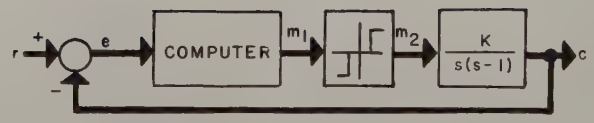


FIG. 11. GENERALIZED OPTIMUM SYSTEM

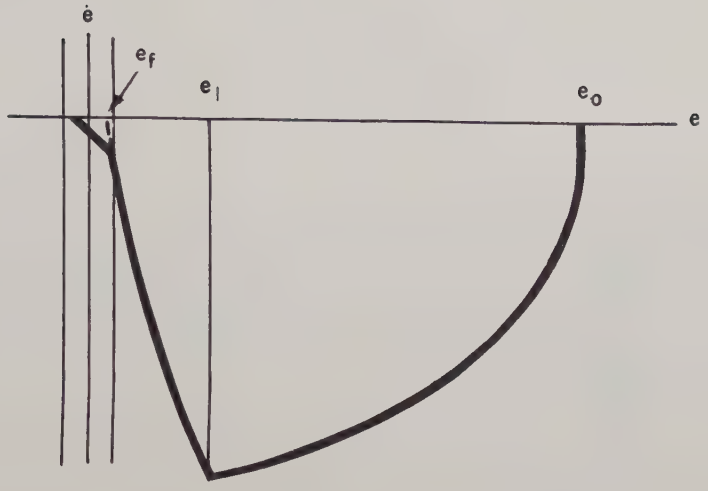


FIG.12. OPTIMUM SYSTEM

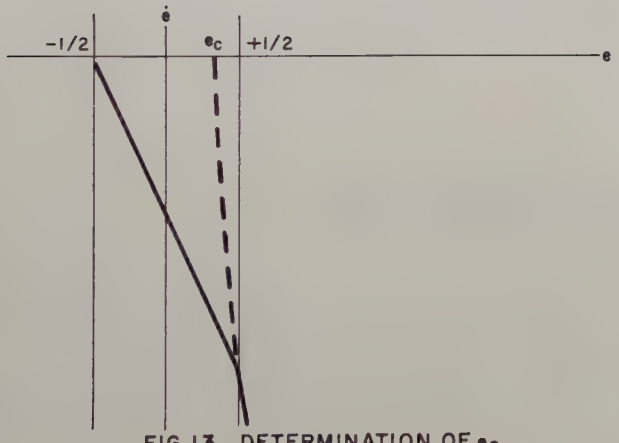


FIG.13. DETERMINATION OF e_c

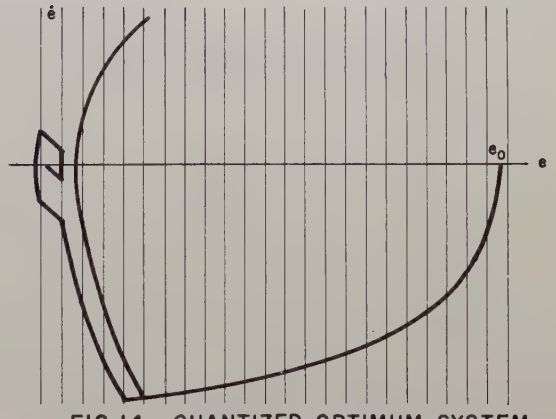


FIG.14 QUANTIZED OPTIMUM SYSTEM
EFFECT OF UNIFORM QUANTIZATION

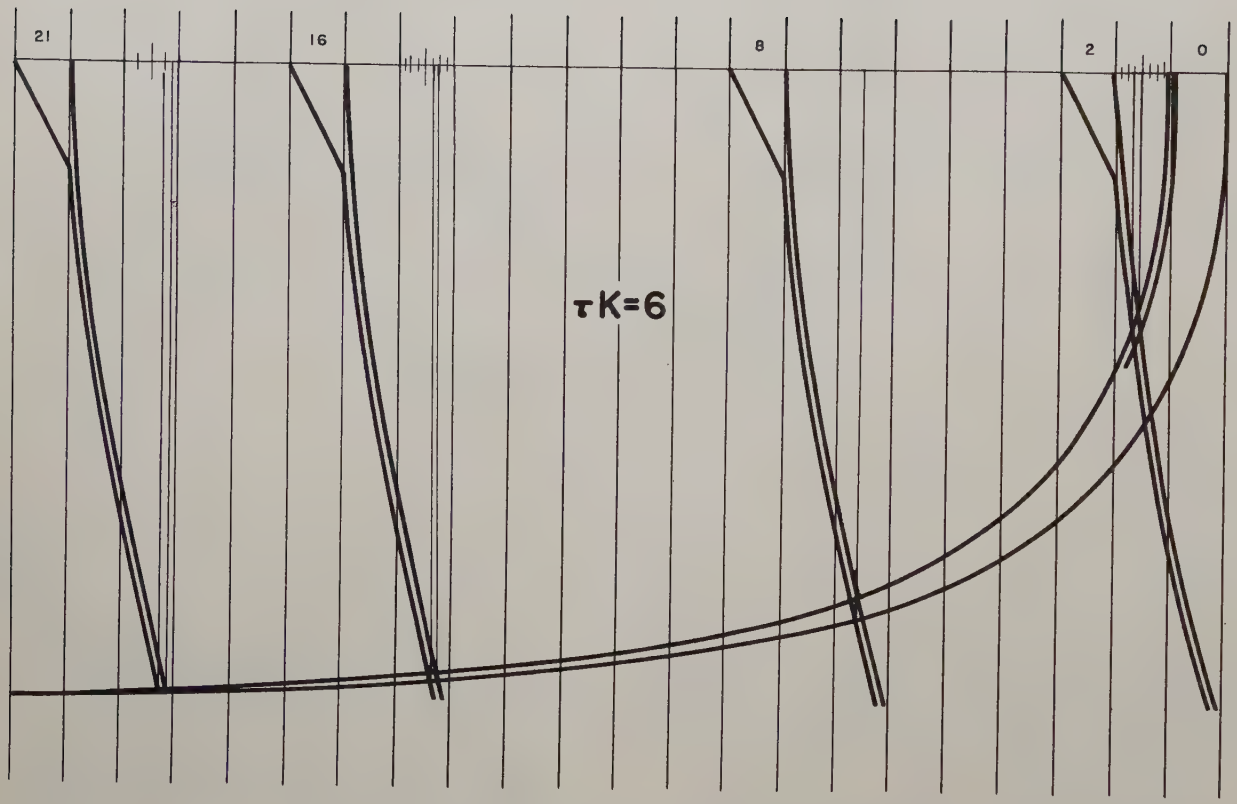


FIG. 15. A DESIGN PROBLEM

SIMPLIFIED METHOD OF DETERMINING TRANSIENT RESPONSE FROM FREQUENCY RESPONSE OF LINEAR NETWORKS AND SYSTEMS

Victor S. Levadi
Antenna Laboratory
Department of Electrical Engineering
The Ohio State University
Columbus 10, Ohio

Summary

Knowing the frequency response of a linear system, a method is presented for obtaining the time response of the system to an impulse, step, or ramp function input, without performing graphical integrations.

The transient response is of the form

$$f(t) = \sum_i A_i G(\omega_i t)$$

where a different function G is used to determine the response to each of the three types of input.

Tables of the functions G(x) are provided.

An example is given to illustrate the simplicity and accuracy of this method. The results are compared with the exact time response.

Introduction

The prediction of the transient performance of linear systems based on their steady state behavior is a problem continually encountered in design and analysis of communication networks and control systems. The mathematical treatment of this problem is quite easily handled by means of Fourier or Laplace transform techniques. However, the transformation of these results from the analytical domain to the realm of numerical reality is often a tedious process at best.

Both Guillemin[1] and Floyd[2] have presented approximate methods of obtaining the impulse response of linear systems. In order to obtain the step or ramp response, these results must be integrated by a numerical or graphical technique.

Guillemin's method, for small values of time, requires taking the difference of large numbers which are nearly equal. This makes the numerical calculation an exacting task.

This discussion will present a method whereby the response of a linear system to an impulse, step or ramp input can be computed directly by a relatively simple means. The problem of computation for small values of time is greatly reduced, and, with the tables provided in the Appendix, the calculations are of a simpler form than for either of the previous methods.

The required information can be obtained experimentally since only a graphical representation, rather than an analytical expression, for the system frequency response is needed.

Analytical Development

Definition of Input Functions

Throughout this discussion the impulse, step, and ramp functions will be defined as follows:

$$\text{Impulse function} \equiv \lim_{\delta \to 0} g(t, \delta)$$

where

$$g(t,\delta) \equiv \begin{cases} 1/\delta, & 0<t<\delta \\ 0, & \text{elsewhere} \end{cases}$$

$$\text{Step function} \equiv \begin{cases} 1, & t>0 \\ 0, & t<0 \end{cases}$$

$$\text{Ramp function} \equiv \begin{cases} t, & t>0 \\ 0, & t<0 \end{cases} .$$

Preliminary Development

Guillemin[3] has shown that only the real, or imaginary, part of the frequency response of a network is sufficient to determine the transient response of that network.

If $f(t)$ and $F(j\omega)$ are a Fourier transform pair, i.e.,

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega)e^{jt\omega}d\omega, \tag{1}$$

and

$$F(j\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt, \tag{2}$$

where

$$f(t) \equiv 0 \quad \text{for } t < 0,$$

then Guillemin shows that

$$f(t) = \frac{2}{\pi}\int_{0}^{\infty} R(\omega)\cos t\omega \, d\omega, \tag{3}$$

and

$$f(t) = \frac{2}{\pi}\int_{0}^{\infty} I(\omega)\sin t\omega \, d\omega. \tag{4}$$

$R(\omega)$ and $I(\omega)$ are the real and imaginary parts of $F(j\omega)$, respectively.

$$F(j\omega) = R(\omega) + jI(\omega). \tag{5}$$

$f(t)$ is the response of the system to an impulse function.

Consider the problem at hand. Given the frequency response function, it is desired to determine the transient response of the network or system.

Since either the real or imaginary part of the frequency response function is sufficient to determine the transient response, we choose to use only the real part, $R(\omega)$.

Physical limitations on the system require that

$$\lim_{j\omega\to\infty} F(j\omega) = 0 . \tag{6}$$

Hence,

$$\lim_{\omega\to\infty} R(\omega) = 0 . \tag{7}$$

If the system is stable

$$\lim_{j\omega\to 0} j\omega F(j\omega) = \lim_{t\to\infty} f(t) . \tag{8}$$

Since the right hand side of (8) is finite,

$$\lim_{\omega\to 0} R(\omega) = \text{constant} . \tag{9}$$

With restrictions (7) and (9) on $R(\omega)$, consider an arbitrary function $R(\omega)$ such as that shown in Fig. 1a.

Response to Impulse Input

In order to perform the integration of (3) the curve $R(\omega)$ is approximated by a series of N straight line segments as shown by dashed lines in Fig. 1a and again in Fig. 1b. For the illustration the approximate curve contains 5 segments. N = 5. The first segment begins at $\omega = 0$ and extends to $\omega = \omega_1$. The coordinates of the end points of the first segment are $(0, R_o)$ and (ω_1, R_1). The slope of this segment is R_1'.

$$R_1' = \frac{R_1 - R_o}{\omega_1} . \tag{10}$$

For the i^{th} segment the beginning and end points are $(\omega_{i-1}, R_{i-1}), (\omega_i, R_i)$.

$$R_i' = \frac{R_i - R_{i-1}}{\omega_i - \omega_{i-1}} . \tag{11}$$

Also

$$R_N \equiv 0, \tag{12}$$

and

$$R_{N+1}' \equiv 0. \tag{13}$$

Integration of (3) by parts gives

$$f(t) = \frac{2}{\pi}\left[R(\omega)\,\frac{\sin t\omega}{t}\,\Bigg|_{\omega=0}^{\omega=\infty} - \int_0^{\infty}\frac{R'(\omega)\sin t\omega}{t}\,d\omega\right]. \tag{14}$$

Conditions (7) and (9) require that the first term of the right hand side of (14) be zero.

Substituting the straightline approximation for $R(\omega)$ into (14) and considering (13), yields

$$f_o(t) = -\frac{2}{\pi}\int_0^{\omega N} \frac{R'(\omega)\sin t\omega}{t}\, d\omega, \qquad (15)$$

where $f_o(t)$ is the impulse response of the system using the straight line approximation for $R(\omega)$.

Breaking the integral (15) into the sum of integrals along each line segment of the approximate $R(\omega)$, and noting that

$$R'(\omega) = R_i' \text{ for } \omega_{i-1} \le \omega \le \omega_i, \qquad (16)$$

there results

$$f_o(t) = -\frac{2}{\pi}\sum_{i=1}^{N}\int_{\omega_{i-1}}^{\omega_i} \frac{R_i'\sin t\omega}{t}\, d\omega. \qquad (17)$$

Integrating and rearranging terms gives

$$f_o(t) = \sum_{i=0}^{N} \frac{2}{\pi}(R_i'-R_{i-1}')\omega_i^2 \frac{\cos \omega_i t}{(\omega_i t)^2}, \qquad (18)$$

where, by definition

$$R_o' \equiv 0 . \qquad (19)$$

Recognizing that

$$\sum_{i=0}^{N}(R_i'-R_{i+1}') = R_o' - R_{N+1}' = 0, \qquad (20)$$

(18) can be rewritten as

$$f_o(t) = \sum_{i=1}^{N} \frac{2}{\pi}(R_i'-R_{i-1}')\omega_i^2\left[\frac{\cos \omega_i t - 1}{(\omega_i t)^2}\right]. \qquad (21)$$

Making the following definitions:

$$b_i = (R_{i+1}' - R_i')\omega_i, \qquad (22)$$

$$G_o(x) = \frac{2}{\pi}\,\frac{1-\cos x}{x^2}, \qquad (23)$$

(21) becomes

$$f_o(t) = \sum_{i=1}^{N} b_i\omega_i G_o(\omega_i t) . \qquad (24)$$

Response to Step Input

The response of the system to a step function is the time integral of the impulse response.

$$f_1(t) = \int_0^t f_o(\tau)\, d\tau \qquad (25)$$

where $f_1(t)$ is the approximate system response to a step input.

Defining

$$G_1(x) \equiv \int_0^x G_o(u)\, du , \qquad (26)$$

(24), (25), and (26) give

$$f_1(t) = \sum_{i=1}^{N} b_i G_1(\omega_i t). \qquad (27)$$

It can easily be verified from (23) and (26) that

$$G_1(x) = \frac{2}{\pi}\left[\mathrm{Si}(x) - \frac{1-\cos x}{x}\right], \qquad (28)$$

where

$$\mathrm{Si}(x) = \int_0^x \frac{\sin u}{u}\, du . \qquad (29)$$

Response to Ramp Input

In a similar manner, the ramp response of the system is the time integral of the step response.

$$f_2(t) = \int_0^t f_1(\tau)\, d\tau, \qquad (30)$$

where $f_2(t)$ is the approximate response to a ramp input. Defining

$$G_2(x) \equiv \int_0^x G_1(u)\, du, \qquad (31)$$

(27), (30) and (31) give

$$f_2(t) = \sum_{i=1}^{N} \frac{b_i}{\omega_i} G_2(\omega_i t), \qquad (32)$$

where

$$G_2(x) = \frac{2}{\pi}\left[x\,Si(x)-(1-\cos x)-Cin x\right], \quad (33)$$

and

$$Cin\ x = \int_0^x \frac{1-\cos u}{u}\,du\ . \quad (34)$$

The Functions G_o, G_1, G_2

The functions G_o, G_1, and G_2 are shown in Fig. 2.

A brief table of these functions is provided in the Appendix.*

The process of computing the time response of the network becomes a relatively simple matter. The constants b_i need only be computed once for a given system.

The only operations required in the computations are multiplication and addition. These calculations provide excellent results with slide rule accuracy. With a desk-type calculator the process of multiplication and summation can be combined in the single process of "accumulative multiplication," further simplifying the computation.

*Copies of a more extensive and detailed table of these functions will be available upon request from the Antenna Laboratory, Department of Electrical Engineering, The Ohio State University, Columbus 10, Ohio.

Error

The two sources of error in this technique are due to approximation of the frequency response curve by straight line segments, and also the error in computation.

Approximation Error. The difference between the true and approximate frequency response functions can be reduced by increasing N, the number of line segments in the approximate function. As N gets very large the approximation error becomes very small.

The amount of computation is proportional to N. Therefore, the choice of the approximate $R(\omega)$, and N is a compromise between accuracy and ease of computation.

Another form of approximation error is apparent when considering the approximate impulse response, $f_o(t)$, for small t.

From (23) it can be shown that

$$\lim_{x\to 0} G_o(x) = \frac{1}{\pi}\ . \quad (35)$$

Using (35) and the limit of (24) as $t \to 0$ gives

$$\lim_{t\to 0} f_o(t) = \frac{1}{\pi}\sum_{i=1}^{N} b_i\omega_i\ . \quad (36)$$

(22) and (36) reduce to

$$\lim_{t\to 0} f_o(t) = \frac{1}{\pi}\sum_{i=1}^{N} (R_{i-1}-R_{i+1})\omega_i. \quad (37)$$

If the approximation to $R(\omega)$ is not such that the right hand side of (37) is 0, then the approximate impulse response, $f_o(t)$, will not begin at 0. Rather, its initial value is that given by (37).

Even if the approximate $f_o(t)$ does not begin at 0, the useful information derived from this curve will not be altered significantly. The information generally desired from the impulse response is peak amplitude, frequency of oscillation, and settling time. It can be seen from the example given later that these quantities derived from the approximate response compare favorably with the true values, even though the approximate curve does not begin at 0.

For the step and ramp response

$$\lim_{x\to 0} G_1(x) = 0, \quad (38)$$

and

$$\lim_{t\to 0} G_2(x) = 0. \qquad (39)$$

Hence, it is obvious that both the computed step and ramp responses will always converge smoothly to 0 as $t \to 0$.

Computational Error. The functions $G_1(x)$ and $G_2(x)$ approach values of 1 and $x - 2/\pi \ln x$ for large x. Hence, the numerical computation of (27) and (32) for large t involves taking the difference of nearly equal numbers. To reduce the effects of error in this process (27) and (32) can be written as

$$f_1(t) = \sum_{i=1}^{N} b_i + \sum_{i=1}^{N} b_i\left[G_1(\omega_i t) - 1\right] \qquad (40)$$

$$f_2(t) = \sum_{i=1}^{N} b_i t + \sum_{i=1}^{N} \frac{b_i}{\omega_i}\left[G_2(\omega_i t) - \omega_i t\right]. \qquad (41)$$

Using (22), (40) and (41) reduce to

$$f_1(t) = R_o + \sum_{i=1}^{N} b_i\left[G_1(\omega_i t) - 1\right] \qquad (42)$$

$$f_2(t) = R_o t + \sum_{i=1}^{N} \frac{b_i}{\omega_i}\left[G_2(\omega_i t) - \omega_i t\right]. \qquad (43)$$

The first terms of (42) or (43) represent a quasi steady state or equilibrium value. This response has the same shape as the input function. The summation terms of (42) and (43) represent the deviation of the response from the input wave form.

For large time the deviation term of (42) approaches 0.

If $R_1' = 0$ the deviation term of (43) approaches the value

$$\frac{2}{\pi} \sum_{i=1}^{N} \frac{b_i}{\omega_i} \cos \omega_i t$$

for large time. Since the deviation terms are much smaller than the equilibrium terms, the calculation of (42) and (43) does not involve small differences of large numbers.

Programming on Digital Computer

Both of the previously mentioned errors can be reduced or eliminated by programming the problem on a digital computer. Obviously, this greatly reduces the computational error.

Since the machine computation is relatively easy and fast, a large number of very short straight line segments can be used to approximate $R(\omega)$. This reduces the approximation error. I In fact, as N gets very large, the results will approach the exact time response.

The functions G_0, G_1, G_2 can be computed with relative ease by using Serracchioli's[5] method for generating the Si and Cin functions.

It should be noted that the computation of the step or ramp response using the G functions involves only one approximation, replacing $R(\omega)$ by the straight line segments. However, if the step and ramp responses were computed by successive numerical integration of the impulse response, the result would contain the cumulative error of each integration.

Summary of Formulae

To summarize the results so far:

	Small t	Large t
Impulse Response $f_o(t)$	$\sum_{i=1}^{N} b_i \omega_i G_o(\omega_i t)$	same
Step Response $f_1(t)$	$\sum_{i=1}^{N} b_i G_1(\omega_i t)$	$R_o + \sum_{i=1}^{N} b_i\left[G_1(\omega_i t) - 1\right]$
Ramp Response $f_2(t)$	$\sum_{i=1}^{N} \frac{b_i}{\omega_i} G_2(\omega_i t)$	$R_o t + \sum_{i=1}^{N} \frac{b_i}{\omega_i}\left[G_2(\omega_i t) - \omega_i t\right]$

where

$$b_i = (R_{i+1}' - R_i')\,\omega_i .$$

Numerical Example

As an example, the approximate impulse and step response for a feedback system represented by the block diagram of Fig. 3 will be computed.

The open loop frequency response function is

$$G(j\omega) = \frac{35.6}{j\omega(.01\,j\omega+1)(.02\,j\omega+1)}$$

The closed loop frequency response function becomes

$$F(j\omega) = \frac{G(j\omega)}{1+G(j\omega)}$$

$$= \frac{1.78\times10^5}{(-150\omega^2+1.78\times10^5)+j(-\omega^3+5000\,\omega)}$$

$R(\omega)$ and the straight line approximation are shown in Fig. 4.

$R(\omega)$ could also have been obtained directly from $G(j\omega)$ by use of Floyd's chart.[4]

The following values are obtained from this figure:

$R_0 = 1$ $\quad N = 4$
$R_1 = 1$ $\quad \omega_1 = 23$
$R_2 = -.65$ $\quad \omega_2 = 41.5$
$R_3 = -.65$ $\quad \omega_3 = 50$
$R_4 = 0$ $\quad \omega_4 = 108$

The solution is now straightforward, as follows:

$$R_1' = \frac{1-1}{26} = 0 \qquad R_3' = 0$$

$$R_2' = \frac{-.65-1}{41.5-23} = -.0892 \qquad R_4' = .013$$

$b_1 = (-.0892-0)(23) = -2.051$ $\quad b_1\omega_1 = -47.18$
$b_2 = 3.701$ $\quad b_2\omega_2 = 153.8$
$b_3 = .560$ $\quad b_3\omega_3 = 28.01$
$b_4 = -1.210$ $\quad b_4\omega_4 = -135.5$

$$f_0(t) = -47.18\,G_0(23t)+153.8\,G_0(41.5t)+28.01\,G_0(50t)-135.5\,G_0(108t).$$

$$f_1(t) = -2.051\,G_1(23t)+3.701\,G_1(41.5t)+.560\,G_1(50t)-1.210\,G_1(108t).$$

For $t = 0.05$

$$f_0(.05) = -47.18\,G_0(1.15)+153.8\,G_0(2.075)+28.01\,G_0(2.5)-135.5\,G_0(5.4)$$
$$= -47.18(.2847)+153.8(.2193)+28.01(.1835)-135.5(.0079)$$
$$= 24.37.$$

$$f_1(.05) = -2.051(.3530)+3.701(.5879)+.560(.6735)-1.210(.9006)$$
$$= .739.$$

The exact time response calculated by the Fourier transform method is

$$f_0(t) = 14.2\,e^{-120.8t} + 44\,e^{-14.6t}\cos(35.5t-109^\circ30');$$

$$f_1(t) = -.1175\,e^{-120.8t} - 1.146\,e^{-14.6t}\cos(35.5t-40^\circ51').$$

The approximate and exact time responses are compared in Figs. 5 and 6.

Conclusion

A method has been presented whereby the impulse, step, or ramp response of a linear system can be computed knowing the frequency response. Any one of the three time responses can be computed separately. No graphical integrations are necessary.

With the table of functions provided, the computation of the transient response is considerably simpler than for previously described methods.

Acknowledgements

The author is grateful for the suggestions and encouragement of Dr. F. C. Weimer, Dr. R. L. Cosgriff, and other staff members of the Antenna Laboratory of The Ohio State University.

Also, appreciation is given to the staff of the Numerical Computation Laboratory of The Ohio State University for their many hours of patient aid in the preparation of the table of functions provided herewith.

References

1. Guillemin, E. A., "Computational techniques which simplify the correlation between steady-state and transient response of filters and other networks," Proc. National Electronics Conference, Vol. IX, pp. 513-532, 1954.

2. Brown, G. S. and Campbell, D. P., Principles of Servomechanisms, pp. 332-350, John Wiley and Sons, Inc., New York, 1948.

3. Guillemin, op. cit.

4. Brown and Campbell, op. cit.

5. Serracchioli, F., "Calculated Values of Self and Mutual Impedances for Parallel Short Dipoles," Report 662-23, 24 March 1959, Antenna Laboratory, The Ohio State University Research Foundation; prepared under Contract DA 36-039 sc 70174, U. S. Army, Signal Corps Engineering Laboratories, Fort Monmouth, New Jersey; pp. 2-4.

Appendix

The following brief table of functions is sufficient for most slide rule calculations. For greater accuracy and ease of computation, more extensive and detailed tables are available from the Antenna Laboratory, Department of Electrical Engineering, The Ohio State University.

x	$G_0(x)$	x	$G_0(x)$	x	$G_0(x)$	x	$G_1(x)$	$G_2(x)$	x	$G_1(x)$	$G_2(x)$
0.0	0.3183	10.0	0.01171	20.0	0.00094	0.1	0.0318	0.0031	6.0	0.9028	3.8629
2	3172	2	1050	2	.00122	2	0636	0073	2	9028	4.0435
4	3141	4	0919	4	150	3	0953	0149	4	9028	2240
6	3089	6	785	6	177	4	1268	0257	6	9029	4046
8	3021	8	652	8	202	5	1581	0398	8	9031	5851
1.0	0.2927	11.0	0.00524	21.0	0.00223	6	1891	0570	7.0	0.9036	4.7658
2	2819	2	404	2	241	7	2198	0773	2	9044	9466
4	2696	4	297	4	254	8	2472	0982	4	9056	5.1275
6	2559	6	204	6	263	9	2801	1271	6	9071	3088
8	2411	8	128	8	265	1.0	0.3096	0.1565	8	9089	4903
2.0	0.2254	12.0	0.00069	22.0	0.00263	1	3387	1888	8.0	0.9110	5.6723
2	2089	2	28	2	256	2	3671	2240	2	9134	8548
4	1920	4	6	4	243	3	3950	2620	4	9161	6.0376
6	1749	6	0	6	227	4	4223	3028	6	9189	2211
8	1586	8	11	8	207	5	4489	3463	8	9218	4051
3.0	1408	13.0	0.00035	23.0	0.00184	6	4749	3925	9.0	0.9248	6.5898
2	1242	2	71	2	160	7	5001	4419	2	9278	7750
4	1083	4	116	4	135	8	5246	4924	4	9307	9609
6	0932	6	168	6	110	9	5483	5460	6	9336	7.1473
8	0790	8	224	8	86	2.0	0.5713	0.6019	8	9362	3342
4.0	0.06579	14.0	0.00280	24.0	0.00064	1	5934	6601	10.0	0.9387	7.5217
2	5378	2	336	2	44	2	6147	7205	2	9409	7096
4	4299	4	387	4	27	3	6352	7829	4	9428	8980
6	3346	6	432	6	15	4	6548	8474	6	9446	8.0867
8	2521	8	469	8	06	5	6736	9138	8	9460	2758
5.0	0.01824	15.0	0.00498	25.0	0.00000	6	6915	9820	11.0	0.9472	8.4651
2	1251	2	516	2	0	7	7086	1.0520	2	9481	6546
4	0797	4	524	4	4	8	7248	1236	4	9488	8442
6	456	6	522	6	10	9	7401	1969	6	9493	9.0341
8	217	8	509	8	20	3.0	0.7546	1.2716	8	9496	2239
6.0	0.00070	16.0	0.00487	26.0	0.00033	1	7683	3477	12	9498	9.4319
2	05	2	456	2	48	2	7811	4251	13	9500	10.364
4	11	4	419	4	64	3	7931	5038	14	9515	11.314
6	73	6	376	6	81	4	8043	5837	15	9555	12.267
8	180	8	329	8	97	5	8148	6646	16	9606	13.225
7.0	0.00320	17.0	0.00281	27.0	0.00113	6	8245	7466	17	9646	14.188
2	481	2	232	2	127	7	8343	8378	18	9662	15.154
4	653	4	185	4	139	8	8417	9132	19	9664	16.120
6	825	6	141	6	149	9	8492	9977	20	0.9668	17.086
8	990	8	101	8	156	4.0	0.8561	2.0830	21	9684	18.054
8.0	0.01139	18.0	0.00067	28.0	0.00159	2	8681	2553	22	9710	19.024
2	1268	2	39	2	160	4	8777	4230	23	9733	19.996
4	1371	4	19	4	157	6	8853	6062	24	9745	20.970
6	1445	6	6	6	152	8	8912	7839	25	9747	21.944
8	1489	8	0	8	143	5.0	0.8955	2.9625	30	0.9795	26.827
9.0	0.01502	19.0	0.00002	29.0	0.00132	2	8986	3.1419	35	9820	31.728
2	1485	2	10	2	120	4	9006	3218	40	9838	36.643
4	1441	4	25	4	105	6	9018	5020	50	9873	46.502
6	1371	6	45	6	090	8	9025	6824	60	9894	56.385
8	1280	8	68	8	75						

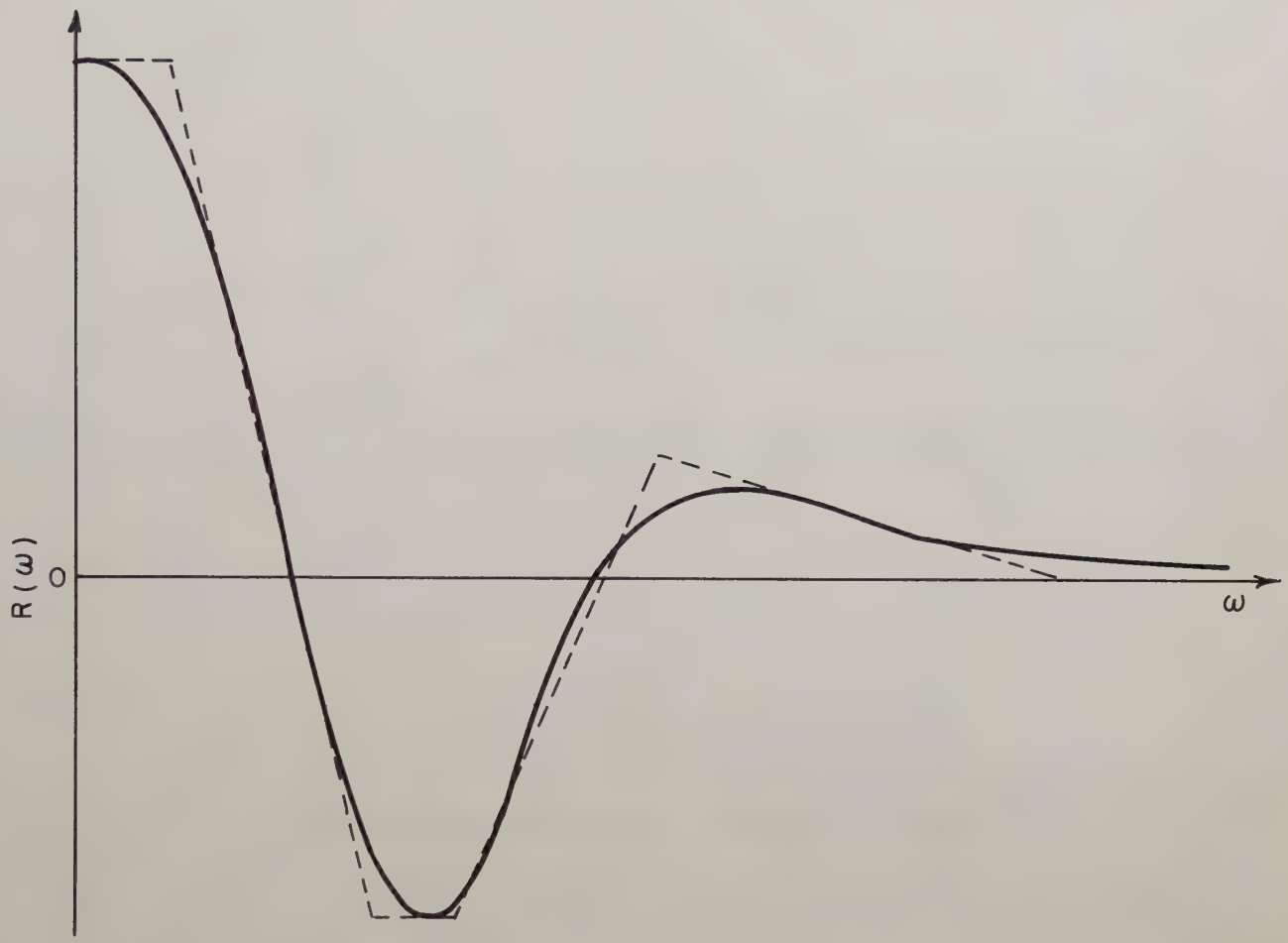


Fig. 1a. Real part of frequency response function and straight line approximation (N = 5).

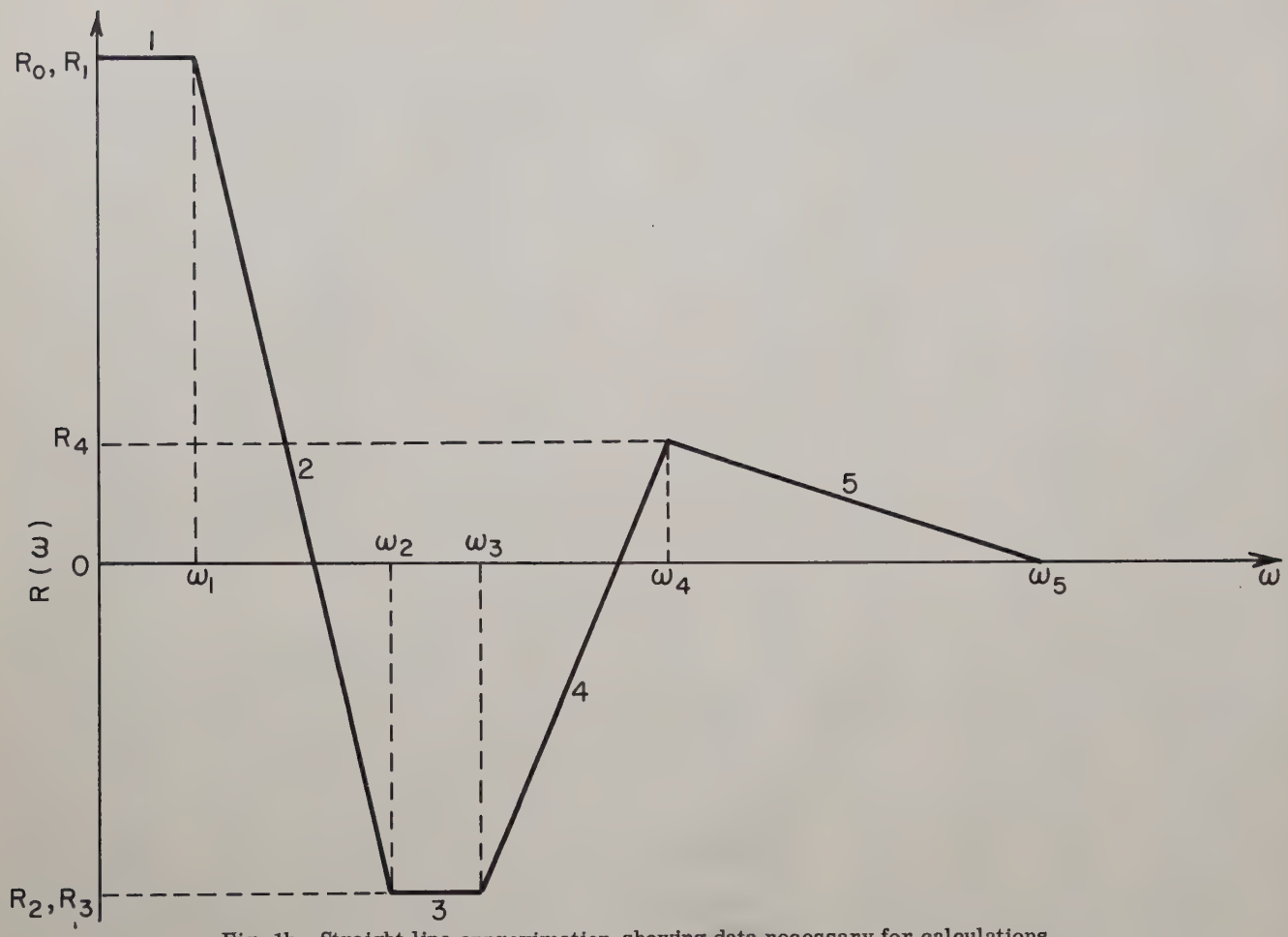


Fig. 1b. Straight line approximation showing data necessary for calculations.

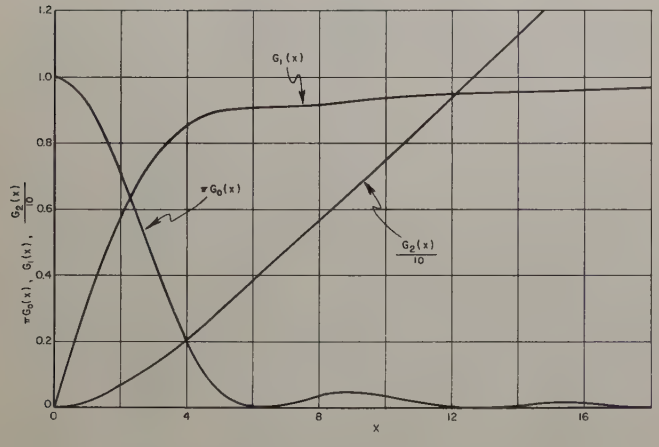


Fig. 2. The functions G_0, G_1, and G_2.

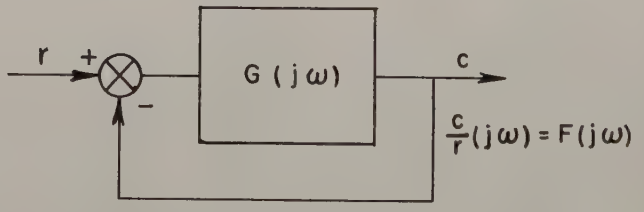


Fig. 3. Block diagram of linear system with unity feedback.

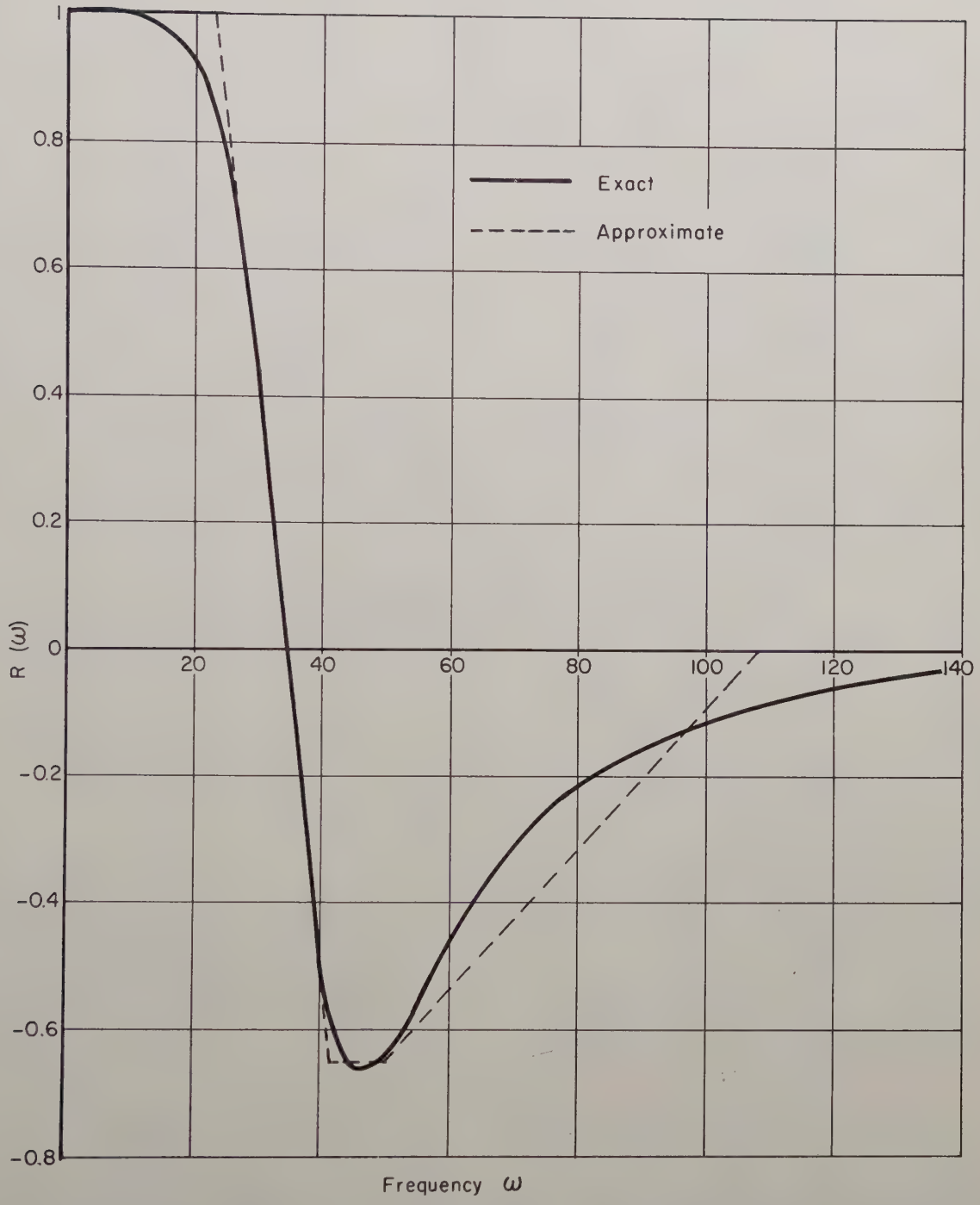


Fig. 4. Real part of frequency response function and straight line approximation for the example.

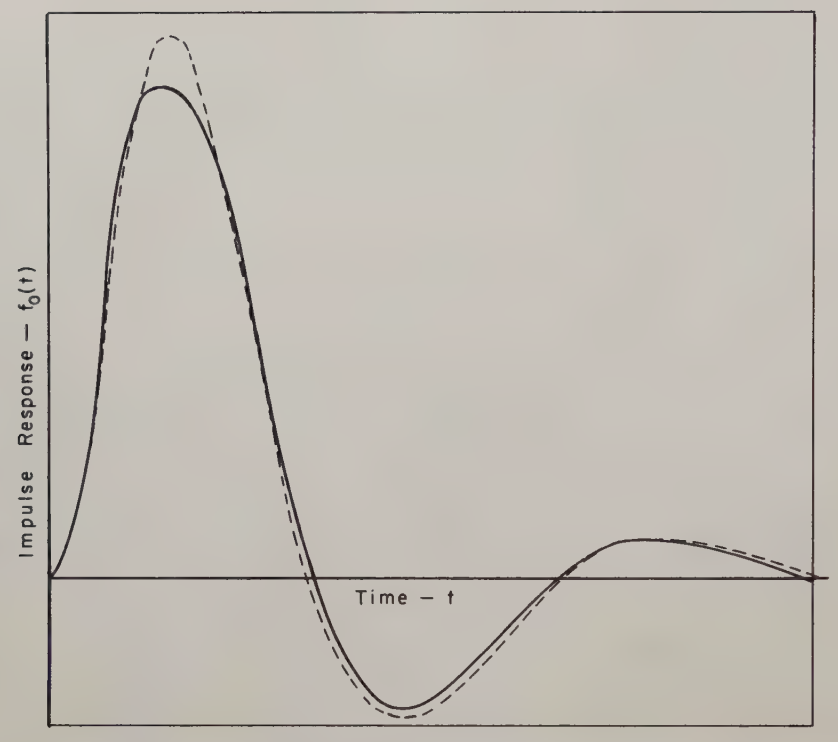


Fig. 5. Comparison of the exact and approximate impulse response.

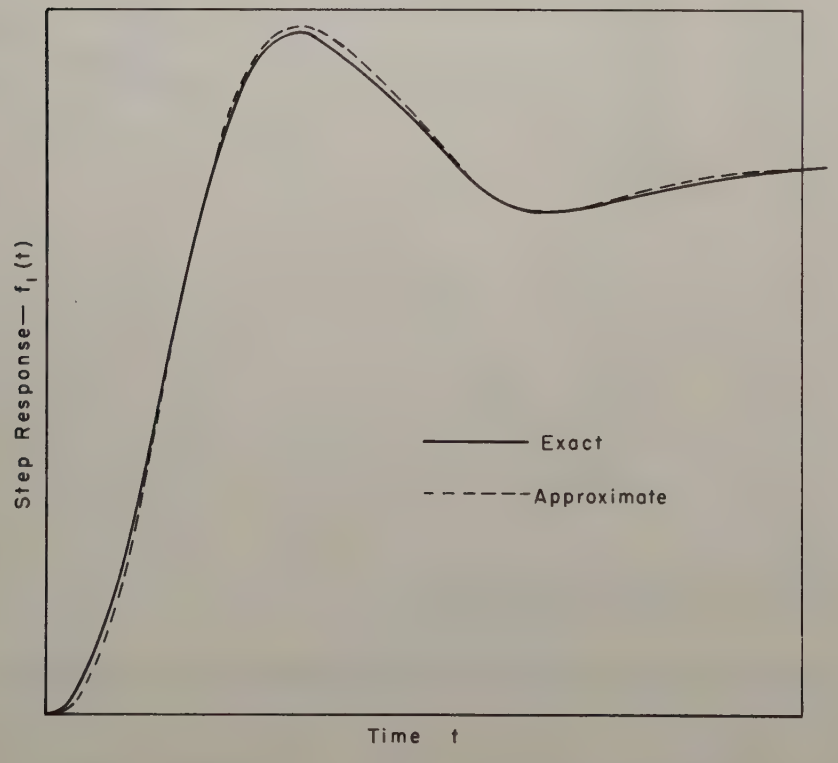


Fig. 6. Comparison of the exact and approximate step response.

A NEW METHOD OF ANALYSIS OF SAMPLED-DATA SYSTEMS

A. Papoulis
Polytechnic Institute of Brooklyn
Brooklyn, New York
and
Burroughs Corporation

Summary

In many sampled-data systems the sampling interval T is "small" and the response $r_s(t)$ closely approximates the response r(t) of the continuous system; one is then interested in evaluating the difference $r_s(t) - r(t)$ for various values of T. In this paper this difference will be given as a power series in T whose coefficients can easily be determined in terms of the continuous response; if one wants to estimate the size of T for $r_s(t)$ to equal r(t) within a specified error, the first term of this expansion will give an adequate measure of the error and hence of the maximum permissible T. Furthermore, since the resulting series converges rapidly, the expansion provides a simple method of evaluating $r_s(t)$ for a given T.

The method is applied to a feedback system with a sampler; the singularities of the p-rational system function that gives the actual response at the sampling points, are obtained by a displacement of the singularities of the continuous system function.

I. The Euler Summation Formula Applied to Sampled-Data

We define the output of a sampler (Fig. 1) by

$$\overset{*}{f}(t) = T \sum_{n=0}^{\infty} f(nT)\, \delta\,(t - nT) \tag{1}$$

The output of a system to f(t) and $\overset{*}{f}$ we denote by r(t) and $r_s(t)$ respectively; we did not use $\overset{*}{r}(t)$ for the output since $r_s(t)$ and r(t) are not of course related by an equation similar to (1). The above definition of $\overset{*}{f}(t)$ differs from the usual by the factor T; this factor was introduced so that

$$\lim_{T \to 0} r_s(t) = r(t) \tag{2}$$

The output of our system to the given input f(t) is a function of the sampling interval T; it can therefore be written as a series

$$r_s(t) = r_0(t) + r_1(t)\, T + r_2(t)\, T^2 + \dots \tag{3}$$

Clearly

$$r_0(t) = r(t) \tag{4}$$

and our problem is to determine the remaining coefficients in (3).

With H(p) the system function and

$$h(t) = \mathcal{L}^{-1}\left\{H(p)\right\} \tag{5}$$

its impulse response, we have for a given

$$t = nT^{-} \tag{6}$$

$$r(t) = \int_0^t f(\tau)\, h\,(t - \tau)\, d\tau = \int_0^t \phi(\tau)\, d\tau \tag{7}$$

where

$$\phi(\tau) = f(\tau)\, h\,(t - \tau) \tag{8}$$

and

$$\begin{aligned} r_s(t) &= \sum_{k=0}^{n-1} T\, f(kT)\, h\,(t - kT) \\ &= \sum_{k=0}^{n-1} T\, \phi(kT) \end{aligned} \tag{9}$$

Thus r(t) is the area under $\phi(\tau)$ and $r_s(t)$ is the area of the inscribed staircase (shaded in Fig. 2). The problem of relating these two areas is very common in numerical quadratures and there are solutions of various forms; a quadrature formula that would lead to a simple determination of r_s should contain no interior points but only the values of ϕ and its derivatives at $\tau = 0$ and $\tau = nT$. The following satisfies these requirements.

Euler Summation Formula

It can be shown (see Appendix) that if F(x) is analytic then

$$\begin{aligned} \int_a^b F(x)\, dx = &\sum_{k=1}^{n-1} T\, F\,(kT) \\ &- \sum_{k=1}^{\infty} \frac{T^k}{k!} B_k \left[F^{(k-1)}(b) - F^{(k-1)}(a) \right] \end{aligned} \tag{10}$$

where $(b-a)/n = T$ and the B_k's are the tabulated Bernoulli numbers:

$$B_1 = -\frac{1}{2},\ B_3 = B_5 = \ldots = B_{2n+1} = \ldots = 0$$

$$B_2 = \frac{1}{6},\ B_4 = -\frac{1}{30},\ B_6 = \frac{1}{42},\ B_8 = -\frac{1}{30}$$

$$B_{10} = \frac{5}{66},\ B_{12} = -\frac{691}{2730},\ B_{14} = \frac{7}{6} \ldots$$

This formula is used to evaluate an integral in terms of the values of F(x) at x = kT, and the values of its derivatives at the end-points; we shall use it in reverse: From (10) we have

$$\sum_{k=0}^{n-1} T\ \phi(kT) = \int_0^t \phi(\tau)\, d\tau + \sum_{k=1}^{\infty} \frac{T^k}{k!} B_k \left[\phi^{(k-1)}(nT) - \phi^{(k-1)}(0) \right] \qquad (11)$$

and with $\phi(\tau)$ as in (8)

$$r_s(t) = r(t) - \frac{T}{2}\left[f(t)h(0) - f(0)h(t) \right] + \ldots \qquad (12)$$

If T is smaller than the smallest time constant in the response, then (12) converges rapidly; it gives a method of evaluating $r_s(t)$. For "small" T

$$r_s(t) = r(t) - \frac{T}{2}\left[f(t)h(0) - f(0)h(t) \right]$$

We thus obtain a simple estimate of T for $r_s(t)$ to equal r(t) within a given error.

Equation (12) could have been obtained from the formula for $r_s(t)$ derived with the usual methods if all terms were expanded into series in T and collected; however the resulting expressions are indeed complicated and it is not so easy to recognize them in the simple form (12).

Example

$$f(t) = u(t) \qquad H(p) = \frac{17}{(p+1)^2 + 4^2}$$

We chose a simple example to facilitate the exact determination of $\bar{r}(t)$; the series in (12) can be as simply evaluated for any H(p). The rate of convergence is the same for any system whose time constant are of the same order of magnitude. From (12) we have for f(t) = u(t)

$$r_s(t) = r(t) + \frac{T}{2}\left[h(t) - h(0) \right] + \frac{T^2}{12}\left[h'(t) - h'(0) \right] + \frac{T^4}{720}\left[h^{(3)}(t) - h^{(3)}(0) \right] + \ldots$$

For our example

$$h(t) = \frac{17}{4} e^{-t} \sin 4t, \qquad r(t) = 1 - e^{-t}\cos 4t - \frac{1}{4} e^{-t} \sin 4t$$

The exact response at the sampling points, evaluated with the usual methods, is given by

$$r_s(t) = A\,(1 - e^{-t}\cos 4t) + B e^{-t} \sin 4t$$

where

$$A = \frac{17}{4}\, T\, \frac{e^{-T}\sin 4T}{1 - 2e^{-T}\cos 4T + e^{-2T}}$$

$$B = \frac{17}{4}\, T\, \frac{e^{-2T} - e^{-T}\cos 4T}{1 - 2e^{-T}\cos 4T + e^{-2T}}$$

In Figures 3, 4 and 5 we have plotted the exact value of $r_s(t)$ and the functions

$$s_1(t) = r(t) + \frac{T}{2}\left[h(t) - h(0) \right]$$

$$s_2(t) = s_1(t) + \frac{T^2}{12}\left[h'(t) - h'(0) \right]$$

for T = .05, .2 and .5.

Even for T = .5 which is larger than $1/\beta = .25$, $s_2(t)$ gives a satisfactory estimate of $r_s(t)$ with an error less than 3%, for T = .05 the term $s_1(t)$ suffices; the error is less than 1%. $s_2(t)$ agrees with $r_s(t)$ up to the 5th significant figure.

Steady State

The steady state $\bar{r}_s(t)$ can be readily evaluated from (12); the terms containing h(t) and all its derivatives tend to zero for large t, hence with $\bar{r}(t)$ the steady state of the continuous system we have

$$\bar{r}_s(t) = \bar{r}(t) - \frac{T}{2} f(t)\, h(0) + \frac{T^2}{12}\left[f'(t)h(0) - f(t)h'(0) \right] + \ldots \qquad (13)$$

Special cases

(a) f(t) = u(t)

$$\bar{r}_s(t) = \bar{r}(t) - \frac{T}{2} h(0) - \frac{T^2}{12} h'(0) - \frac{T^4}{720} h^{(3)}(0) + \ldots$$

(b) f(t) = sin ωt

with $\bar{r}_s(t) = A_s \sin \omega t + B_s \cos \omega t$

$$\bar{r}(t) = A \sin \omega t + B \cos \omega t$$

we have from (13)

$$A_s = A - \frac{T}{2} h(0) - \frac{T^2}{12} h'(0) + \ldots$$

$$B_s = B + \frac{T^2}{12} \omega h(0) + \ldots$$

Hold Circuit

The hold circuit can be considered as part of the system; the combined system function is given by

$$H_1(p) = \frac{1 - e^{-pT}}{pT} H(p)$$

and its impulse response by

$$h_1(t) = \frac{1}{T} \left[a(t) - a(t-T) \right] \tag{14}$$

where $a(t)$ is the step response of the original system function; therefore (see (12)).

$$r_s(t) = r(t) - \frac{1}{2} \left[f(t)a(0) - f(0)a(t) - f(0)a(t-T) \right] + \ldots \tag{15}$$

since $a(t) = 0$ for $t<0$. The steady state is given by

$$\bar{r}_s(t) = \bar{r}(t) - \frac{1}{2} f(t)\, a(0) + \frac{T}{12} \left[f'(t)a(0) - f(t)a'(0) \right] + \ldots \tag{16}$$

as we can readily see from (13) and (14).

II. Transforms of Sampled-Data

The Euler formula can be used to determine the transform $\overset{*}{F}(p)$ of $\overset{*}{f}(t)$ as a series in T. With

$$F(p) = \int_0^\infty e^{-pt} f(t)dt = \int_0^\infty q(t)dt \tag{17}$$

where

$$q(t) = e^{-pt} f(t) \tag{18}$$

and $\overset{*}{f}(t)$ as in (1), we have

$$\overset{*}{F}(p) = \int_0^\infty e^{-pt} \overset{*}{f}(t)dt = \sum_{n=0}^{\infty} Tq(nT) \tag{19}$$

therefore (see (10))

$$\overset{*}{F}(p) = F(p) - \sum_{n=1}^{\infty} T^n \frac{B_n}{n!} q^{(n-1)}(0) = F(p) + \frac{T}{2} f(0) - \frac{T^2}{12} \left[f'(0) - pf(0) \right] + \ldots \tag{20}$$

since $q^n(\infty) = 0$ for every n.

Response

With $R(p) = H(p)\, F(p)$,

the transform $\overset{*}{R}_s(p)$ of $\overset{*}{r}_s(t)$ is given by

$$\overset{*}{R}_s(p) = \overset{*}{H}(p)\, \overset{*}{F}(p) \tag{21}$$

as one can readily see from the definition. $\overset{*}{R}_s(p)$ can be written in the form

$$\overset{*}{R}_s(p) = \left[R_0(p) + R_1(p)\, T + \ldots \right]^* \tag{22}$$

where $R_0(p) = R(p)$ is the transform of the continuous response, and

$$R_k(p) = \mathcal{L} \left\{ r_k(t) \right\} \qquad k = 1, 2, \ldots$$

are functions to be determined. From (20), (21) and (22) we obtain

$$\left[R + R_1\, T + \ldots \right] + \frac{T}{2} \left[r(0) + T\, r_1(0) + \ldots \right] + \ldots = \left[H + \frac{T}{2} h(0) + \ldots \right] \left[F + \frac{T}{2} f(0) + \ldots \right] \tag{23}$$

Equating equal powers of T we obtain R_1, R_2, ... ; the result agrees with (12) but the details are omitted.

Example

$$f(t) = u(t), \qquad H(p) = \frac{17}{(p+1)^2 + 4^2}$$

$$r(t) = 1 - e^{-t} \cos 4t - \frac{1}{4} e^{-t} \sin 4t$$

$$R_1(p) = \frac{1}{2} r(0) + \frac{1}{2} h(0)\, F(p) + \frac{1}{2} f(0)\, H(p) = \frac{1}{2} H(p)$$

hence

$$r_1(t) = \frac{1}{2} h(t)$$

$$R_2(p) = \frac{1}{12} \left[pH(p) - 17\, F(p) \right]$$

$$r_2(t) = \frac{1}{12} \left[h'(t) - h'(0) \right]$$

The results agree with the previous development of the same example.

III. Feedback

Consider the system of Fig. 6; its system function is given by

$$\overset{*}{H}_f = \frac{\overset{*}{H}}{1 + k\overset{*}{H}} \tag{24}$$

For a system without the sampler

$$H_f = \frac{H}{1 + kH} \quad \text{where} \quad H(p) = \frac{N(p)}{D(p)} \tag{25}$$

The poles of H_f we denote by $p_1, \ldots p_m \ldots, p_n \ldots$, thus

$$D(p_m) + kN(p_m) = 0 \quad m = 1, \ldots, n \tag{26}$$

The poles p'_m of $\overset{*}{H}_f$ depend on the sampling interval T; for small T they can be obtained from p_m by a linear displacement

$$\Delta p_m = p'_m - p_m$$

To determine Δp_m we expand H as in (20) and insert into (24)

$$\overset{*}{H}_f = \frac{H(p) + \frac{T}{2}h(0) - \frac{T^2}{12}\left[h'(0) - ph(0)\right] + \ldots}{1 + k\,H(p) + \frac{kT}{2}h(0) - \frac{kT^2}{12}\left[h'(0) - ph(0)\right] + \ldots} \tag{27}$$

p'_m is a root of

$$D + kN + \left[\frac{kT}{2}\,h(0) - \ldots\right] D = 0 \tag{28}$$

We shall retain only first order effects.

p_m simple: if $h(0) \neq 0$ then we can readily obtain from (28) the following expression for Δp_m.

$$\Delta p_m = -\frac{kT}{2}\,h(0)\,A_m \tag{29}$$

where

$$A_m = \left.\frac{D(p)\,(p-p_m)}{D + kN}\right|_{p=p_m} \tag{30}$$

is the residue of $1/(1 + kH)$ at $p = p_m$; i.e., the displacement is proportional to T. If $h(0) = 0$ we consider the next term in (28); we then obtain

$$\Delta p_m = \frac{kT^2}{12}\,A_m\,h'(0) \tag{31}$$

where A_m as in (30).

p_m multiple: if the multiplicity of p_m is r then (28) gives the following results. For $h(0) \neq 0$

$$\Delta p_m = \sqrt[r]{-\frac{kT}{2}\,B_m H(0)} \tag{32}$$

where

$$B_m = \left.\frac{D(p)\,(p-p_m)^r}{D + kN}\right|_{p=p_r} \tag{33}$$

Thus from each pole of H_f of multiplicity r there result r poles for $\overset{*}{H}_f$ on the vertices of a regular polygon with p_m as center. For $h(0) = 0$

$$\Delta p_m = \sqrt[r]{\frac{k\,T^2}{12}\,h'(0)B_m} \tag{34}$$

where B_m as in (33).

Examples

1. $H(p) = \dfrac{1}{p^2 + ap + b}$ $\quad h(0)=0,\ h'(0)=1$

$$D + kN = p^2 + ap + b + k, \qquad \begin{matrix} p_1 \\ p_2 \end{matrix} = -a \pm j\beta$$

For $\beta \neq 0$

$$A_1 = -\frac{k}{2j\beta} \qquad \Delta p_1 = \frac{k^2T^2}{24\beta}\,j$$

For $\beta = 0$

$$\beta_1 = -k \qquad p_1 = \pm\frac{-k^2T^2}{12} = \pm\frac{kT}{2\sqrt{3}}\,j$$

(see Fig. 7a)

2. $H(p) = \dfrac{p + c}{p^2 + ap + b}$ $\quad h(0) = 1$

$$D + kN = p^2 + ap + b + k(p + c) \qquad \begin{matrix} p_1 \\ p_2 \end{matrix} = -a \pm j\beta$$

For $\beta \neq 0$

$$A_1 = \frac{-k(p_1 + c)}{2j\beta} \qquad \Delta p_1 = \frac{k^2T}{4\beta}\,\frac{p_1 + c}{j}$$

For $\beta = 0$

$$B_1 = -k\,(c - a) \qquad \Delta p_1 = \pm\sqrt{\frac{k^2T}{2}\,(c-a)}$$

(see Fig. 7b)

Appendix

To prove (10) it suffices to take n = 1 since, if the formula is established for each interval and the results are added, the terms $F^{(k)}(x)$ evaluated in the interior points cancel. Expanding $F(x)$, $F'(x)$, ... into a power series and integrating from 0 to T we obtain

$$\int_0^T F(x)dx = F(0)T + \frac{F'(0)T^2}{2!} + \frac{F''(0)T^3}{3!} + \ldots \qquad 1$$

$$\int_0^T F'(x)dx = F'(0)T + \frac{F''(0)T^2}{2!} + \ldots \qquad C_1$$

$$\vdots$$

$$\int_0^T F^{(k)}(x)dx = F^{(k)}(0)\,T + \frac{F^{(k+1)}(0)T^2}{2!} + \ldots \qquad C_k$$

We next multiply the k+1 equation by C_k k=1, 2, ... and add; the constants C_k are so chosen that the coefficients of $F^{(k)}(0)$ k=1, 2, ... in the r.h. side of the resulting equation are all equal zero. For this to be true the C_k's must satisfy the following equations.

$$\frac{T^2}{2} + C_1 T = 0 \qquad \text{hence } C_1 = -\frac{T}{2}$$

$$\frac{T^3}{3!} + C_1\frac{T^2}{2!} + C_2 T = 0 \qquad \text{hence } C_2 = \frac{T^2}{12}$$

$$\frac{T^4}{4!} + C_1\frac{T^3}{3!} + C_2\frac{T^2}{2!} + C_3 T = 0 \quad \text{hence } C_3 = 0$$

$$\frac{T^{n+1}}{(n+1)!} + C_1\frac{T^n}{n!} + C_2\frac{T^{n-1}}{(n-1)!} + \ldots + C_n T = 0$$

From the above equation (10) readily follows and the Bernoulli numbers are obtained through a recursion formula.

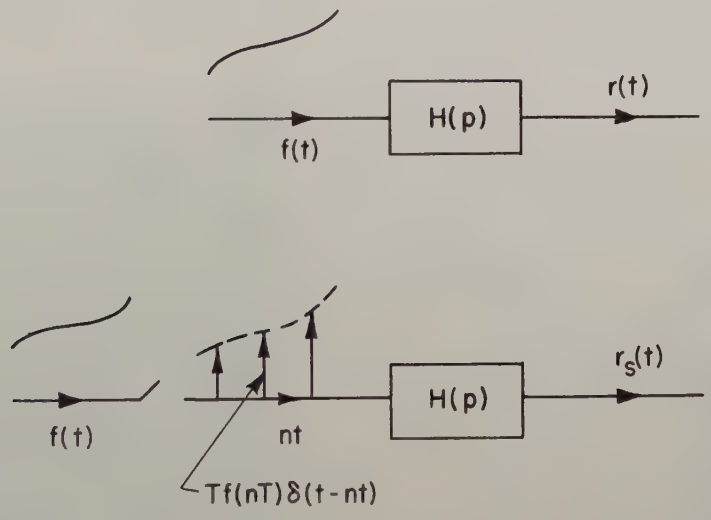
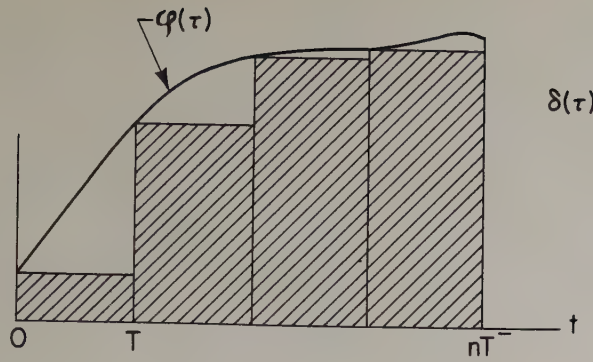


Fig. 1.



$$r(nT^-) = \int_0^{nT^-} \varphi(\tau)\,d\tau$$

$$r_s(nT^-) = \sum_{K=0}^{n-1} T\varphi(KT)$$

Fig. 2.

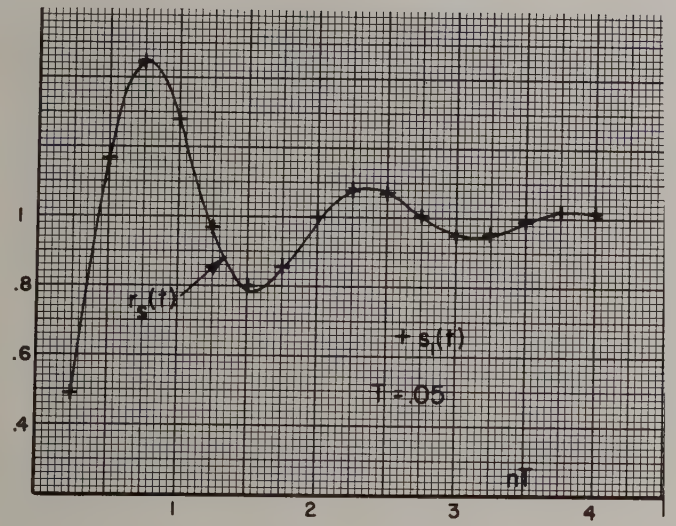


Fig. 3.

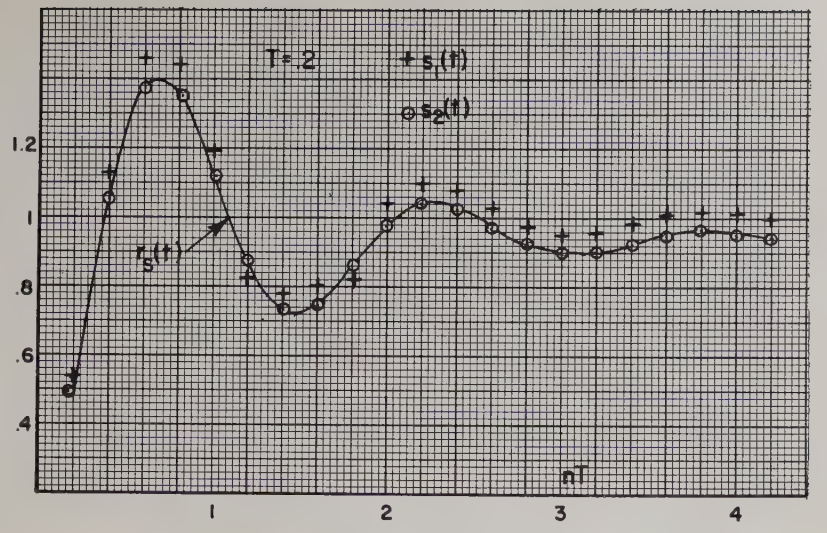


Fig. 4.

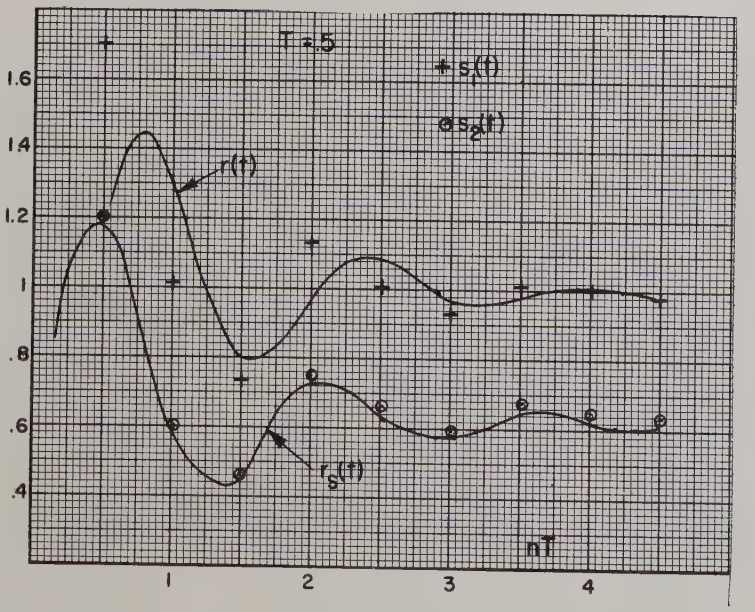


Fig. 5.

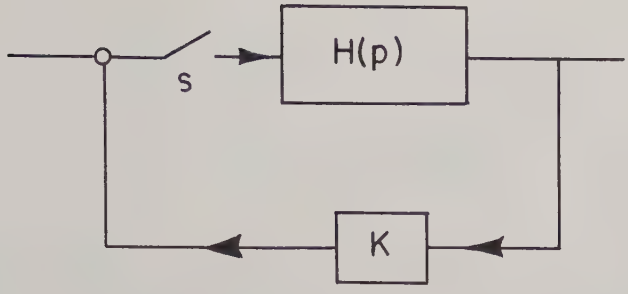


Fig. 6.

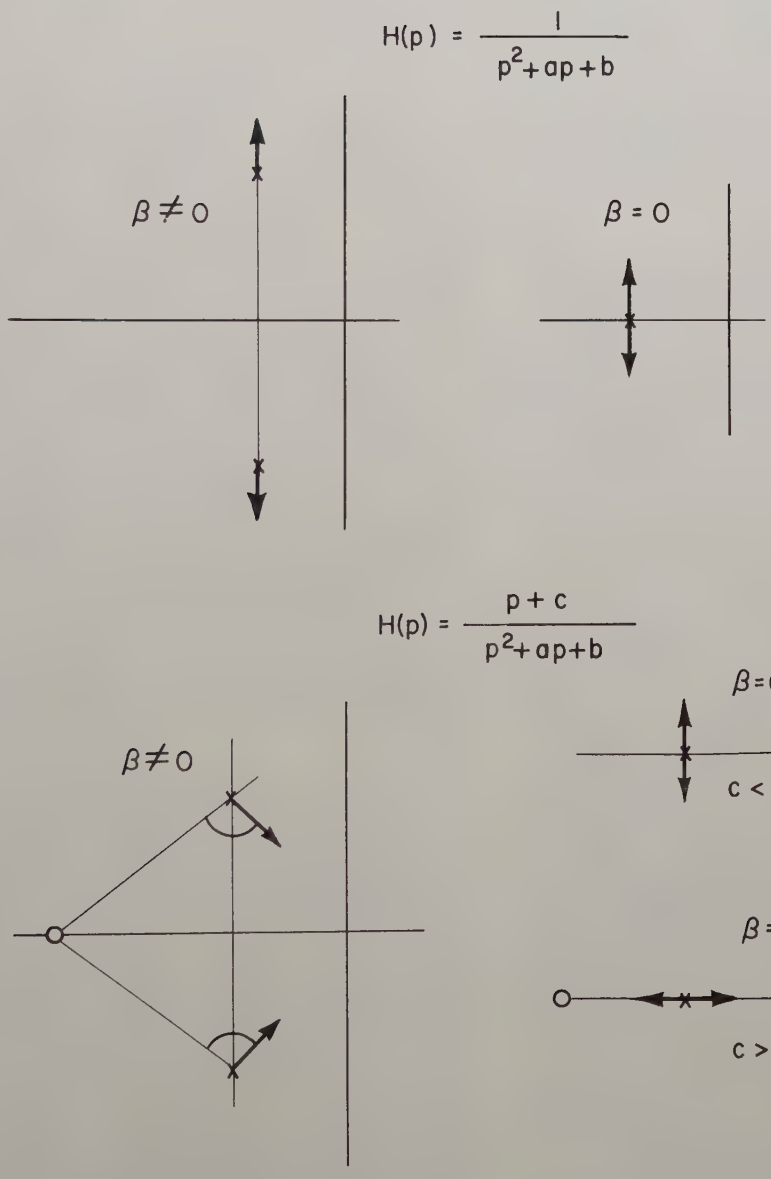


Fig. 7.

STATISTICAL FILTER THEORY FOR TIME-VARYING SYSTEMS

Elwood C. Stewart and Gerald L. Smith

National Aeronautics and Space Administration
Ames Research Center
Moffett Field, Calif.

Summary

The guidance and control accuracy of most aerodynamic and space vehicles is limited primarily by the noise disturbances which enter the guidance system. Present theory for the minimization of these noise effects does not take into account the forced kinematic time variations, such as is due to time-varying range, which occur in most guidance problems. Such time variations occur in interplanetary flight, satellite rendezvous, interception of missiles or bombers, and so forth. In this paper an analytical approach is presented for the optimization of systems which are forced both to be time varying and to operate with inputs contaminated with noise.

Two objectives are the establishment of the theoretical optimum performance and a method of synthesizing the optimum control system. Effects of restrictions on the capability of the output element in addition to the noise and the forced time variation are considered. Although an exact analytical solution of the problem does not appear feasible, it is shown how approximate solutions utilizing time-varying control systems can be found. The method is illustrated by a hypothetical example of a homing missile attacking a bomber.

Introduction

The use of statistical filter theory in the design of systems is now well established. It has found application in communication systems, automatic control systems, and weapons systems. The reason for the usefulness of the statistical approach is not hard to see. In the first place the inputs to systems often can be represented only in statistical terms, as, for example, the signal input in a communication system, the motions of targets, etc. Secondly, there are invariably introduced into the system some contaminating signals called noise which are also best handled statistically. Consequently, statistical theory and its application to system design has grown in importance.

There are a great many possible statistical approaches to take. Foremost of these is the well-known Wiener filter theory.[1] Many later works have dealt with a number of modifications to this original work. There are several assumptions which are generally common to many of these theories. One of the principal assumptions is that the systems to which the theory is to be applied are time invariant or at least are not forced to be time varying. Although many problems fall in this category, there is a large class of very important control system problems which are inherently time varying so that present theory is not applicable. Consequently, the purpose of this paper will be to show how statistical theory can be extended to these time-varying problems. In particular, we will be concerned with the synthesis of optimum time-varying systems. More detail than can be given in this paper will be found in reference 2.

Problem Description

In the optimization of time-varying systems it is important to distinguish between two time-varying aspects. The over-all system may be time varying because (1) certain forced time variations appear in the system as a result of its mode of operation, or (2) a time-varying control system is employed purposely because of advantages to be gained by so doing. Both kinds of time variation may occur of course in the same problem.

One of the most important types of forced time variations, and the one with which we will be concerned here, is illustrated in Fig. 1. On the left is a space vehicle in the mid-course guidance of an interplanetary flight, and on the right is a homing missile intercepting an enemy bomber. The guidance of such vehicles is accomplished by employing line of sight information which therefore involves range. Since this range varies continuously during a flight, these problems are necessarily time varying. The time-varying situations illustrated here may be recognized as belonging to a large and important class. Similar situations arise for instance in studies of fire-control systems, aircraft landing systems, and so forth.

The class of time-varying problems illustrated in Fig. 1 can generally be described in terms of linear time-varying differential equations, or, as is completely equivalent, in terms of a block diagram such as shown in Fig. 2. Here we see that the error signal is modified by a time-varying element before it is available for use by the controller. It is this element which establishes the essential time-varying nature of the problem. For the type of problem which was illustrated in Fig. 1 this variation is a kinematic relationship of the form $1/R(t)$, where R is the range between vehicle and destination. There are other time-varying problems in which the forced time variation may occur in some other location in the block diagram or in which more than one time-varying element may occur. However, all such problems may be treated in a similar manner. We will confine our study specifically to the time-varying problem illustrated in Fig. 2.

The two inputs to the system in Fig. 2 are the true signal information and a contaminating noise signal which enters with the desired information. The noise signal is due to the uncertainty in measurement of the signal; it might arise from errors in star tracking, errors in radio communication, or from errors in a radar or an infra-red detection system. These two inputs are modified by the time-varying element. The remaining portion of the block diagram is the control system which closes the kinematic loop. Since the control system is invariably limited in its capabilities, it is generally necessary to impose a constraint on the most critical quantity within the system. Thus, it is desirable that the control system be split up into a controller and output element as shown so that the quantity to be constrained, r, appears explicitly. This quantity might represent a linear or angular acceleration, the control motion of a missile, or the thrust from reaction controls, for example. The output element is the fixed element which might represent the dynamics of the space vehicle, missile, or whatever, depending on the application. Merely for simplicity here, this element is taken to be of a typical form k/s^2. The other element of the control system is the controller which the designer is presumably free to choose.

The optimization problem we wish to solve can now be expressed as follows: If (in Fig. 2) we are given the forced time-varying element, the statistical properties of the signal and the noise, we would like to find the controller which will minimize the error ϵ and yet not require of the constrained quantity r more than the system capabilities. The problem is similar in concept but basically more complicated than the time invariant case which was treated in reference 3. The general class of problem we are concerned with is characterized by the fact that the error need be minimized only at a particular time T (the time of arrival at the destination, target, etc.) but the system capabilities should not be exceeded at any time t_2 previous to T. Ensemble averages are particularly meaningful in this type problem. Mathematically such problems are treated by minimizing the quantity

$$\overline{\epsilon^2(\tau)} + \rho \overline{r^2(t_2)} \qquad (1)$$

where the quantity ρ is a Lagrangian multiplier. From physical reasoning, the restriction on the quantity r should be constant at the maximum permissible value throughout the entire flight.

Solution

There are several steps involved in the solution of the problem as stated and they will be enumerated in the following discussion.

Adjoint Theory

The first step is to convert the differential equations in the real-time domain to a new kind of time called adjoint time. The way in which such a concept comes about is as follows. It is known that when a time-varying system is activated at time zero and thereafter subjected to a stationary random process with a constant spectral density of unit magnitude, the mean-square ensemble average of the error is given by

$$\overline{\epsilon^2(t_2)} = 2\pi \int_0^{t_2} h^2(t_2,t_1)dt_1 \qquad (2)$$

where h is the error response at time t_2 due to an impulse introduced at time t_1. For example, in Fig. 3, if an impulse is introduced into the system at time t_1', the error response as shown might be obtained. The value of h at the desired time t_2 is indicated. If the impulse were introduced at another time t_1'' as shown a different response would be obtained and so forth. Note that these responses are plotted as a function of time t for a fixed time t_1. However, in equation (2) it is necessary to have h as a function of t_1 for a fixed t since the integration is with respect to t_1. Such a response could be obtained by cross-plotting the curves given, and the result might appear as given at the bottom of Fig. 3. However, this procedure is completely unsuited to the synthesis problem we have here.

A better procedure is based on a transformation of the real-time differential equations to the corresponding adjoint differential equations.[4] It can be shown that the solution of these equations gives the desired response $h(t_2,t_1)$ as a function of t_1 as required in equation (2). In other words the solution of the adjoint equation is the response shown at the bottom of Fig. 3. The transformation involved here is related to the reciprocity principle which is common in many fields. Two familiar examples are the well-known reciprocity theorem in circuit analysis, and the Rayleigh-Carson antenna law. Here reciprocity manifests itself by an interchange of the input and output positionwise and also timewise, the latter being achieved by running time backwards.

Derivation of Equations

The second step in the solution is to derive the necessary equations, and to examine possible methods of solving these equations. Expressions for the two quantities of interest, the error and restricted quantity should be written in terms of the adjoints of the unknown controller which we wish to determine. This is done by transforming to the adjoint equations or corresponding adjoint block diagram, from which the error and restricted quantity can be written:[2]

Error:

$$\overbrace{\epsilon^2(\tau)} = \int_0^\infty \Bigg(\overbrace{\left\{ \int_0^\infty h(\tau - x) h_N(x)\,dx \right\}^2}^{\text{noise component}} + \overbrace{\left\{ \int_0^\infty [u_o(\tau - x) - h(\tau - x)] h_S(x)\,dx \right\}^2}^{\text{signal component}} \Bigg) d\tau \tag{3}$$

$$h(\tau) = \frac{k}{\tau} \iint c(\tau)\,d\tau\,d\tau - \frac{k}{\tau} \int_0^\tau h(\xi) \iint c(\tau,\xi)\,d\tau\,d\tau\,d\xi \tag{4}$$

Restricted quantity

$$\overbrace{r^2(t_2)} = \int_0^\infty \Bigg(\overbrace{\left\{ \int_0^\infty g(\tau - x, \tau_1) h_N(x)\,dx \right\}^2}^{\text{noise compoment}} + \overbrace{\left\{ \int_0^\infty [u_o(\tau - x - \tau_1) - g(\tau - x, \tau_1)] h_S(x)\,dx \right\}^2}^{\text{signal component}} \Bigg) d\tau \tag{5}$$

$$g(\tau,\tau_1) = \frac{c(\tau,\tau_1)}{\tau} - \frac{k}{\tau} \int_0^\tau g(\xi,\tau_1) \iint c(\tau,\xi)\,d\tau\,d\tau\,d\xi \tag{6}$$

As can be seen, to relate the error to the controller impulse response $c(\tau,\xi)$ requires two equations. The first relates the error to the intermediate but unknown impulse response $h(\tau)$; the remaining factors in this equation are known when the inputs are specified. The second is an integral equation in which this same response h is related to the controller response $c(\tau,\xi)$. Together these two equations relate the error to the controller. A similar situation exists for the two equations for the restricted quantity. Consequently, the quantity to be minimized, $\overbrace{\epsilon^2(\tau)} + \rho \overbrace{r^2(t_2)}$, is expressed in terms of the unknown controller by means of these four equations. It is these equations then together with equation (1) which comprise the exact formulation of the optimization problem.

The problem now is one of solving these equations for the optimum controller $c(\tau,\xi)$. Unfortunately these equations are so formidable that exact minimization does not appear feasible. Consequently an approximate solution has been sought.

A clue to such a solution can be found when the restriction is eliminated, that is, the system is assumed to be linear regardless of the magnitude of signals within the system. For this case $\rho = 0$ and the equations can be readily solved. In order to illustrate the solution, a typical example situation for a homing missile attacking an enemy bomber has been taken. For this situation the minimum obtainable error at the interception time, that is, the miss distance, has been determined and is given by the lower dashed curve in Fig. 4 as a function of one of the most important parameters, the noise magnitude. As indicated, this result is valid for all time-varying systems. The important clue is that this curve is identical with that which is obtained for a completely constant-coefficient system as is indicated by the lower solid curve. This result is valid of course only for the case of no restrictions. It does suggest, however, that even with a restriction included, the optimum performances of the time-variant and time-invariant systems ought to be the same. The optimum performmance of time-invariant systems can be determined by well-known optimization methods and, by the above argument, this result should also be the optimum performance of the time-varying system. This is illustrated in Fig. 4 by the two curves with restrictions; the upper solid curve represents the exact solution according to constant-coefficient theory.

There are many compelling reasons to support the above argument. Merely from physical considerations we would expect that as long as the output element and the restriction on the input to this element is the same for both systems, there would be no inherent reason why this output element could not be controlled as well for both systems. There are many other reasons too lengthy for discussion here. Nevertheless we can say that as a good approximation the optimum error performances, with restrictions, of both the time-varying and non-time-varying systems are identical.

Synthesis for Optimum Error Performance

The third step in the solution is to synthesize the controller so that the optimum error performance will be achieved, while the restriction requirement is temporarily ignored. It can be seen from the first error equation, (3), that to have the same error performance as the constant-coefficient system it is only necessary that the response h be the same. This h is obtained along with the optimum error performance from the time-invariant optimization theory. Having h, the second error equation, (4), can then, in principle at least, be solved for the desired controller $c(\tau,\xi)$. However, solutions of this integral equation are not easy to obtain. The difficulty appears to be in the manner of representing the control system by a single impulse

response. Since the impulse response of even a physically very simple time-varying system may be completely unwieldy, we would expect the impulse response for the optimum controller to be even more unmanageable. Consequently, it has been necessary to find another representation.

A more suitable representation which has been found is based on splitting up the controller into several parts, each with its own impulse response. Such a form is illustrated in Fig. 5 where the controller has been broken up into three parts: a time-varying multiplying part and two non-time-varying parts. This form is quite general and yet a practical one easy to construct. Moreover, it is this formulation which is needed to enable the optimum control system to be synthesized. An equation which replaces equation (4) can now be derived - one that relates the response h not to c as in equation (4), but rather to the control system components as it has been broken up in Fig. 5. It can be shown[2] that this equation is

$$h(\tau) = \frac{1}{\tau}\mathcal{L}^{-1}\left\{Y_2(s)\mathcal{L}\left[f(\tau)\mathcal{L}^{-1}\left(\frac{kY_1(s)}{s^2}\right)\right]\right\} - \frac{1}{\tau}\mathcal{L}^{-1}\left\{Y_2(s)\mathcal{L}\left[f(\tau)\int_0^\tau h(\tau - x)\iint ky_1(x)dx\,dx\,dx\right]\right\} \qquad (7)$$

Now for certain time-varying functions $f(\tau)$ this equation can be solved for the remainder of the controller $Y_1(s)$ and $Y_2(s)$. Thus the desired impulse response h and therefore the desired error performance can be achieved. Actually it is found that there are many solutions of equation (7) so that a whole class of time-varying controllers can be generated all of which have the desired error performance.

Satisfying Restriction Requirement

The fourth and last step is to satisfy the desired restriction requirement which has been temporarily ignored. This can be done as follows. Each of the systems which satisfy the desired error performance will have different restriction requirements throughout the interval of interest 0 to T. Thus from this whole class of systems satisfying the error performance, the one or more having the desired restriction characteristics can be chosen. We would expect that if a system with uniform restriction characteristic were chosen, this level would be identical to that imposed in the time-invariant optimization.

Example

To illustrate this method of solution let us consider in more detail the example case of a homing missile attacking an enemy bomber, as was illustrated in Fig. 1. This problem can of course be cast in the block diagram form shown in Fig. 5. The signal would represent the true target position, and the noise would represent the error in the measurement of this target position by the radar; the block $Y_2(s)$ would represent a radar tracking system and filtering and $Y_1(s)$ the autopilot. As has been indicated, there are many controllers which can be synthesized to satisfy the desired error criterion. It will be most instructive to illustrate the results for only two of these controllers. For the first, let us take the simplest form of control, a constant-coefficient controller, for which $f(t)$ in Fig. 5 is a constant. For the second, we will take a simple time-varying controller in which $f(t)$ is range, that is, a range multiplication controller. When the necessary equations are solved, it is found that the first system turns out to be what is known as proportional navigation. Furthermore it is interesting to note that for both systems the form of the required radar tracking system $Y_2(s)$ in Fig. 5 is such as to call for differentiation (as well as filtering) of the line of sight. This result is gratifying because it is exactly what one expects on the basis of heuristic reasoning.

Now let us examine the performance of these two systems. The miss distance (or error) performance of both systems is the same and is given by the upper dotted curve in Fig. 4. However, the demands on acceleration are vastly different as is illustrated in Fig. 6. Also shown in this figure is the design value which represents the desired restriction level on the amount of acceleration. The choice between the two systems is clear. The first, proportional navigation, calls for too little acceleration early in the flight and as a consequence attempts to make up for this deficiency by calling for more than its capabilities later in the flight. Consequently, limiting would occur and the miss distance would be greater than that shown in Fig. 4. The second system, the range multiplication controller, is much better in this regard because the desired restriction level is approximated very closely throughout the entire flight. Thus this system achieves the optimum error performance and satisfies the desired restriction on acceleration very closely. It is therefore an approximate solution to the optimization problem.

Concluding Remarks

It is desirable to point out that although the example presented here was a specific homing missile problem many problems can be cast in a similar form. The reason is that both noise effects and the time-varying range which exists between vehicle and destination are common ingredients in the automatic control and guidance of many vehicles. There are applications, however, which may differ principally in the location at which the noise is introduced. Such changes are not fundamental and similar methods may be used.

It is also important to emphasize that optimization theory does not eliminate all problems in the design of automatic control systems. There are many factors which might influence the specific design of a system as for example, the necessity of using certain fixed and unalterable

elements in the system or the introduction of artificial damping to the vehicle. Fortunately the optimization theory presented herein does not prescribe one unique system; in other words, there are many ways in which the optimum performance can be achieved. For this reason it is usually possible to include other requirements in the design without sacrificing the optimum performance.

References

1. Wiener, Norbert: The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications. The Technology Press, M.I.T., 1949.

2. Stewart, Elwood C., and Smith, Gerald L.: The Synthesis of Optimum Homing Missile Guidance Systems With Statistical Inputs. NASA MEMO 2-13-59A, 1959.

3. Stewart, Elwood C.: Application of Statistical Theory to Beam-Rider Guidance in the Presence of Noise. II - Modified Wiener Filter Theory. NACA TN 4278, 1958.

4. Laning, J. Halcombe, Jr., and Battin, Richard H.: Random Processes in Automatic Control. McGraw Hill Book Co., 1956.

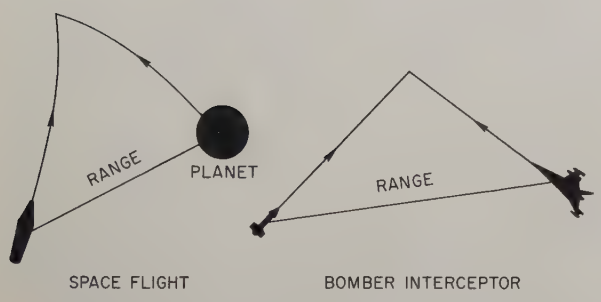


Fig. 1. Geometry of time-varying problem.

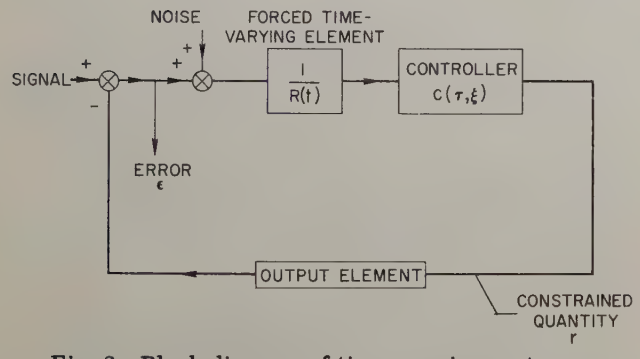


Fig. 2. Block diagram of time-varying system.

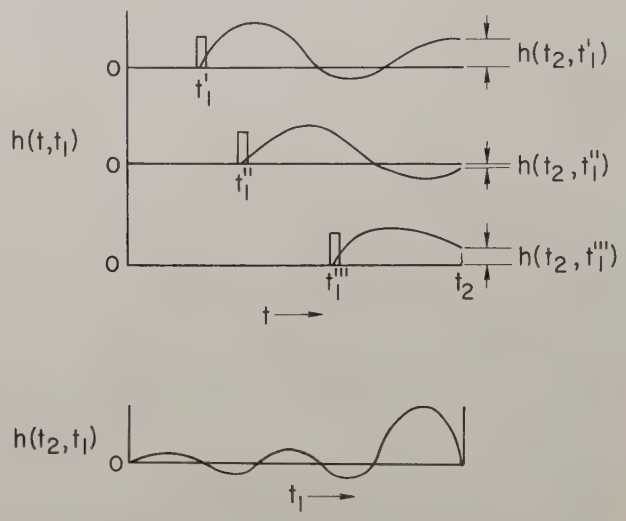


Fig. 3. Evaluation of mean-square ensemble average.

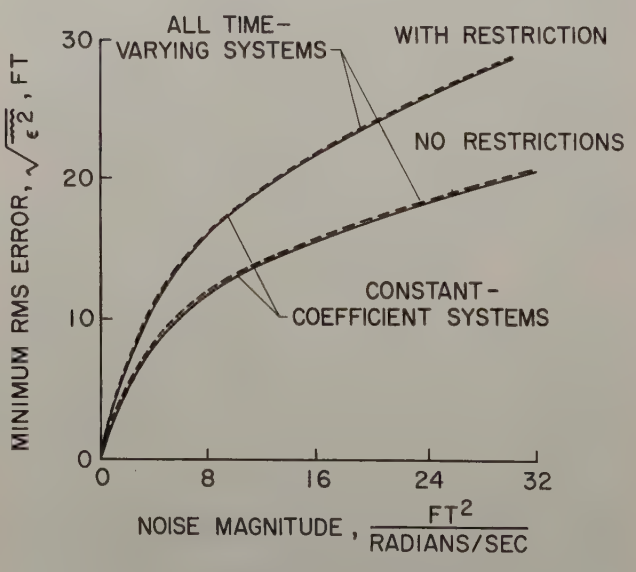


Fig. 4. Optimum error performance.

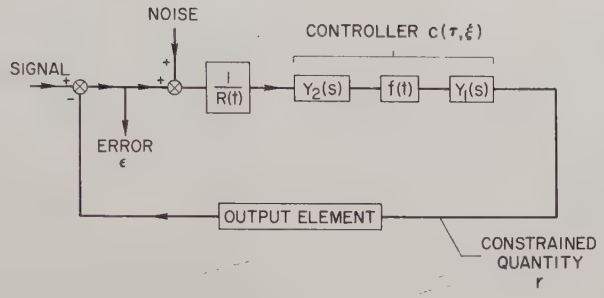


Fig. 5. Block diagram of time-varying system.

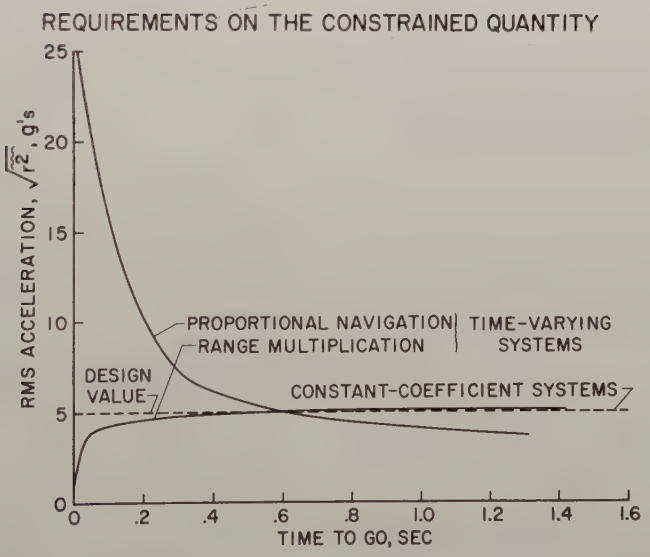


Fig. 6. Requirements on the constrained quantity, acceleration.

ON THE PHASE PLANE ANALYSIS OF NON-LINEAR TIME-VARYING SYSTEMS

Richard F. Whitbeck
Systems Requirements Department
Cornell Aeronautical Laboratory, Inc.
Buffalo 21, New York

Summary

A phase plane technique, which takes advantage of convenient relationships in other planes (for example, displacement versus time) to effect graphical solutions for a non-linear time varying second order differential equation, is developed. Several special cases of this general second order equation are considered. The special case of most practical importance occurs when the differential equation is permitted to become piecewise linear. To demonstrate the simplicity of the technique, for the piecewise linear case, an example involving saturation in an inertially damped position servomechanism is given.

Introduction

In this paper, the phrase "phase plane technique" is interpreted to mean any approximate graphical solution method that interprets a first order derivative as a slope in its appropriate plane (for example, dv/dx defines the slope of a solution curve in the v,x plane, while dx/dt defines the slope of a solution curve in the x, t plane).* Admitting the above as a reasonable definition of a phase plane technique leads rather naturally to the idea of using more than one plane simultaneously to effect approximate solutions to problems described by non-linear time varying differential equations.

The paper deals specifically with a second order equation, although the technique described is general enough to be applied to higher order systems (this is demonstrated in an appendix for a third order non-linear time-varying differential equation).

The accuracy of the technique can be greatly enhanced by carrying along a solution curve in a seemingly "redundant" plane. For example, if solution curves are being constructed simultaneously in the v,x and v,t planes, a construction in the x,t plane should yield the same value of x as was obtained in the v,x plane, and the same value of t as was obtained in the v,t plane. The tangential segments which approximate the solution curves must be decreased if the values in the x,t plane do not check, within some acceptable degree of accuracy, with those obtained in the other two planes.

Several special cases of the general second order differential equation are considered in order that the reader may clearly comprehend the widely varying amounts of work required in effecting a solution as the equation is permitted to take on more restricted meanings.

An important practical objective of the paper will be to introduce the reader to a more generalized Lienard's construction, since it handles the majority of non-linear problems encountered in everyday engineering in a more efficient manner than the majority of other contemporary techniques. An example involving saturation in an inertially damped position servomechanism is given to demonstrate the simplicity of the technique when applied to piecewise linear problems.

The various cases will be considered in the sections to follow, beginning in Section 1 with the general second order differential equation.

Section 1
A General Second Order Equation

This paper will be concerned with the phase plane solution for the second order differential equation

$$a(x,v,t)\,\ddot{x} + b(x,v,t)\,\dot{x} + C(x,v,t)\,x = f(t). \quad (1\text{-}1)$$

In Equation (1-1) $a(x,v,t)$, $b(x,v,t)$ and $c(x,v,t)$ are coefficients dependent on the signal, its derivative and time. $f(t)$ represents the forcing function which, in general, is time dependent.

Letting $\dot{x} = v$, noting that $\ddot{x} = v\,dv/dx$ and dividing through by the coefficient of $\ddot{x}$ gives Equation (1-2).

$$v\frac{dv}{dx} + \frac{b(x,v,t)}{a(x,v,t)}\,v + \frac{C(x,v,t)}{a(x,v,t)}\,x = \frac{f(t)}{a(x,v,t)} \quad (1\text{-}2)$$

or

$$\frac{dv}{dx} = \frac{-\frac{b(x,v,t)}{a(x,v,t)}\,v - \frac{C(x,v,t)}{a(x,v,t)}\,x + \frac{f(t)}{a(x,v,t)}}{v}. \quad (1\text{-}3)$$

For the convenience of discussion let the numerator of Equation (1-3) be $-g(x,v,t)$.

$$\therefore \frac{dv}{dx} = \frac{-g(x,v,t)}{v}. \quad (1\text{-}4)$$

Since $v\frac{dv}{dx} = \frac{dv}{dt}$, Equation (1-4) can be written in a slightly different form:

---

* *An explicit definition of the term "phase plane" does not seem to exist in the literature. Apparently, a "phase plane" has some dependent variable as the absicca, with the derivative of the dependent variable as the ordinate. Hence, velocity, displacement and acceleration, velocity planes would be phase planes, while displacement, time or acceleration, displacement planes would not be.*

$$\frac{dv}{dt} = -g(x, v, t) \quad . \tag{1-5}$$

Another relationship at our disposal is that

$$\frac{dx}{dt} = v \, . \tag{1-6}$$

Equation (1-4) defines the slope of a solution curve in the v , x plane, Equation (1-5) defines the slope of a curve in the v , t plane, and Equation (1-6) defines the slope of a curve in the x , t plane. Notice that between any two of these planes all the variables encountered in Equation (1-1) are present. Given the initial conditions, a tangential approximation to the solution curve can be made in any two of these planes (or all three if so desired).

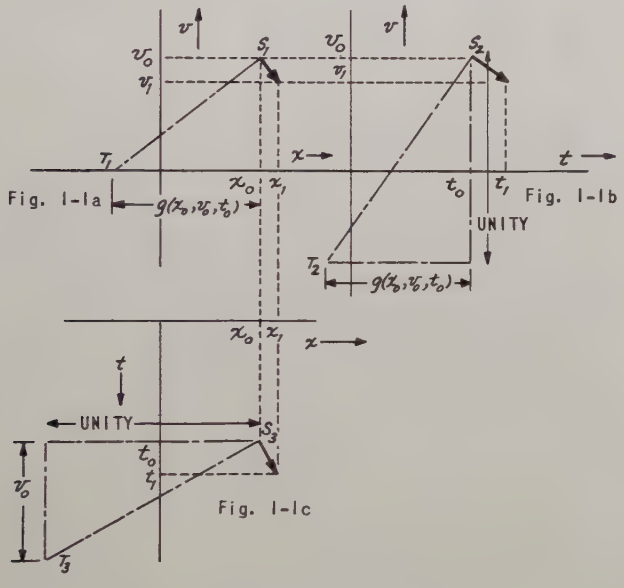


Fig. 1-1. Construction technique.

From Figure 1-1 the basic scheme for approximating the solution curves in the various planes becomes clear. Consider Figure 1-1a. At $t = t_0$, the numerator of Equation (1-4) might be evaluated as $-g(x_0 , v_0 , t_0)$. In the v,x plane a perpendicular dropped from the point x_0 , v_0 to the x axis automatically gives a line segment whose length equals v_0 . A length on the x axis equal to $g(x_0 , v_0 , t_0)$ can be measured off from the point x_0 . The slope of the line $S_1 T_1$ is then:

$$m_{S_1 T_1} = \frac{v_0}{g(x_0, v_0, t_0)} \tag{1-7}$$

The slope of the perpendicular to the line $S_1 T_1$ at v_0 , x_0 is the negative reciprocal of the slope $S_1 T_1$.

$$-\frac{1}{m_{S_1 T_1}} = \frac{-g(x_0, v_0, t_0)}{v_0} \tag{1-8}$$

This slope satisfies Equation (1-4) and hence a short segment of the tangent constitutes a solution in the neighborhood of x_0 and v_0 . The end point of the short tangent segment will give a new value of x and v, say x_1 and v_1 . However a new value of t corresponding to x_1 and v_1 is required before the numerator of Equation (1-4) can be evaluated and the process stated above repeated.

Since x_0 , t_0 and v_0 are known, and $-g(x , v , t)$ has already been evaluated as $-g(x_0 , v_0 , t_0)$, Equation (1-5) can be used to obtain a tangential approximation to the solution curve in the v , t plane. Refer to Figure 1-1b to see how this is done. At the point v_0 and t_0 , a perpendicular is dropped and a unit distance measured off. On a line parallel to the t axis the distance $g(x_0 , v_0 , t_0)$ is measured off, which gives the slope of the line $S_2 T_2$ as:

$$m_{S_2 T_2} = \frac{1}{g(x_0, v_0, t_0)} \quad . \tag{1-9}$$

Therefore the slope of the perpendicular to the line $S_2 T_2$ at v_0 and t_0 is:

$$-\frac{1}{m_{S_2 T_2}} = -g(x_0 , v_0 , t_0) \quad . \tag{1-10}$$

This satisfies Equation (1-5) and hence a short segment of the tangent approximates the solution curve in the neighborhood of v_0 and t_0 .

In the x , t plane the story is much the same (refer to Figure 1-1c). At the point x_0 and t_0 a line parallel to the x axis is drawn and a unit distance measured off. On a line parallel to the t axis the distance v_0 is measured off. The slope of the line $T_3 S_3$ is:

$$m_{T_3 S_3} = -\frac{1}{v_0} \quad . \tag{1-11}$$

The slope of the perpendicular to $T_3 S_3$ at x_0 , t_0 is

$$-\frac{1}{m_{T_3 S_3}} = v_0 \tag{1-12}$$

This satisfies Equation (1-6) and hence a short segment of the tangent approximates the solution curve in the neighborhood of x_0 and t_0 .

By drawing parallel lines between the three planes one can determine the new values of x , v and t to be used in the next approximation to the solution curves. The procedure described above is then repeated as many times as necessary to obtain complete approximations to the solution curves in each plane.

Note that only two planes are actually needed, however carrying along solutions in all three planes enables one to make judicious choices for the lengths of the tangent segments which approximate the solution curves. For example, if the procedure yields a new value x_1 , in the v , x plane and t_1 in the v , t plane, a construction in the x , t plane should yield the same values of

x and t . If the same values are not obtained, then the lengths of the tangent segments must be decreased until an acceptable degree of accuracy is achieved. Hence a self-checking feature is incorporated in the technique by the use of the seemingly redundant third plane.

The technique described above is an extremely simple one, but nevertheless tedious - since a numerical calculation of $g(x, v, t)$ must be made for each tangential approximation to the solution curve. Indeed, we may conclude that the technique constitutes a numerical solution of Equation (1-1)* which uses the various planes to tabulate the results in an accumulative fashion.

In the sections which follow various special cases, which do not require as much numerical calculation, of Equation (1-1) will be considered. The special case of most practical importance to the engineer will occur when Equation (1-1) is permitted to become "piecewise linear", with $f(t)$ equal to a constant.

Section 2
Coefficients Independent of Time

If the coefficients of Equation (1-1) are not explicit functions of time, the equation simplifies to

$$a(x,v)\ddot{x} + b(x,v)\dot{x} + c(x,v)x = f(t). \qquad (2\text{-}1)$$

Using the same change of variable employed in Section 1 gives:

$$\frac{dv}{dx} = \frac{-\frac{b(x,v)}{a(x,v)}v - \frac{c(x,v)}{a(x,v)}x + \frac{f(t)}{a(x,v)}}{v} = \frac{-g(x,v,t)}{v} \qquad (2\text{-}2)$$

Since the numerator of Equation (2-2) is still some function of time, no real simplification results. Of course, the numerical computation of $g(x, v, t)$ may be considerably easier.

The same sort of situation arises when $f(t)$ equals a constant, but the coefficients remain explicit functions of time.

Section 3
Coefficients Independent of Time, Forcing Function Constant

When the coefficients are not explicit functions of time, and $f(t)$ equals a constant (k), then Equation (3-1) results:

$$\frac{dv}{dx} = \frac{-\frac{b(x,v)}{a(x,v)}v - \frac{c(x,v)}{a(x,v)}x + \frac{k}{a(x,v)}}{v} = \frac{-h(x,v)}{v} \qquad (3\text{-}1)$$

** In this paper we have restricted ourselves to a very general second order equation. It is not intended to imply that only second order systems can be treated by this technique of employing convenient relationships between a number of planes. Relationships between higher order planes exist which permit the graphical solution technique discussed here to be applied to higher order systems. An illustration of this is given in Appendix A for a third order system.*

For this special case the work required is materially reduced since only the x , v plane is necessary to construct a complete solution curve (solutions in the v, t and x, t planes can be carried along for their information content). Notice that a numerical computation of $h(x,v)$ is still necessary for each approximation to the solution curve.

Section 4
Damping Velocity Dependent, Spring Force Displacement Dependent

If $a(x,v) \rightarrow a$, $b(x,v) \rightarrow b(v)$ and $c(x,v) \rightarrow c(x)$ Equation (3-1) reduces to

$$\frac{dv}{dx} = \frac{-\frac{b(v)}{a}v - \frac{C(x)}{a}x + \frac{k}{a}}{v} \qquad (4\text{-}1)$$

or

$$\frac{dv}{dx} = \frac{-\phi_{(v)} - \theta_{(x)} + k_1}{v} \qquad (4\text{-}2)$$

For this special case almost all the tedious numerical work can be eliminated. A curve of $\phi(v)$ vs v * can be drawn as can a curve of $\theta(x)$ vs x .* Therefore, at any particular value of x and v (say x_o and v_o), the values $\phi_{(v_o)}$ and $\theta_{(x_o)}$ can be picked from their respective curves with a compass and used to construct the tangential approximation as shown in Figure 4-1.

The slope of the line $S_4 T_4$ is:

$$m_{S_4T_4} = \frac{v_o}{\phi_{(v_o)} + \theta_{(x_o)} - k_1} \qquad (4\text{-}3)$$

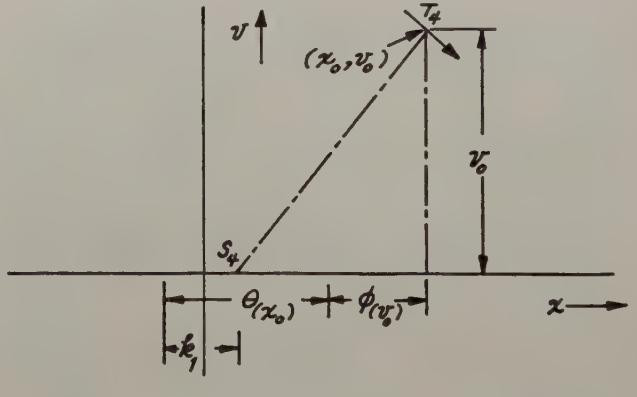


Fig. 4-1. Construction details.

** $\phi(v)$ and $\theta(x)$ must be plotted to the same relative scales as the x , v plane.*

The negative reciprocal is:

$$-\frac{1}{m_{S_4T_4}} = \frac{-\phi_{(v_0)} - \theta_{(x_0)} + k_1}{v_0} \qquad (4\text{-}4)$$

Equation (4-4) satisfies Equation (4-2), hence the construction is valid.

For convenience, one may include the constant k_1 in either of the auxillary curves (for example, plot $[\theta_{(x)} - k_1]$ vs. x) and eliminate one compass measurement.

Section 5
Lienard's Construction

When $a(x,v) \rightarrow a$, $b(x,v) \rightarrow b(v)$ and $c(x,v) \rightarrow c$, Equation (3-1) can be manipulated to read as

$$a\ddot{x} + b(v)\dot{x} + cx = k \quad . \qquad (5\text{-}1)$$

Equation (5-1) is a special case adequately covered by the discussion of Section 4 - however it is interesting to re-examine this equation using a slightly different approach. Divide Equation (5-1) by c and use the change of variable

$$\tau = \sqrt{c/a}\ t \qquad (5\text{-}2)$$

$$\therefore \frac{d\tau}{dt} = \sqrt{c/a} \qquad (5\text{-}3)$$

and

$$\frac{dx}{dt} = \frac{dx}{d\tau} \cdot \frac{d\tau}{dt} = \sqrt{c/a}\ \frac{dx}{d\tau} \qquad (5\text{-}4)$$

also

$$\frac{d^2x}{dt^2} = \frac{dv}{dt} = \frac{dv}{d\tau} \cdot \frac{d\tau}{dt} = \sqrt{c/a}\ \frac{dv}{d\tau} \qquad (5\text{-}5)$$

but

$$\frac{dv}{d\tau} = \frac{d}{d\tau}\left(\frac{dx}{dt}\right) = \sqrt{c/a}\ \frac{d^2x}{d\tau^2} \qquad (5\text{-}6)$$

$$\therefore \frac{d^2x}{dt^2} = c/a\ \frac{d^2x}{d\tau^2} \qquad (5\text{-}7)$$

Substituting Equations (5-4) and (5-7) into Equation (5-1) gives

$$\frac{d^2x}{d\tau^2} + \frac{b(v)}{\sqrt{ac}}\ \frac{dx}{d\tau} + x = \frac{k}{c} = k_2 \qquad (5\text{-}8)$$

Now, let $dx/d\tau = v$, which yields

$$\frac{dv}{dx} = \frac{-\frac{b(v)}{\sqrt{ac}}\ v - x + k_2}{v} \qquad (5\text{-}9)$$

as the phase plane equation. For convenience, let

$$\frac{b(v)}{\sqrt{ac}}\ v - k_2 = \psi_{(v)} \ , \qquad (5\text{-}10)$$

giving

$$\frac{dv}{dx} = \frac{-\psi_{(v)} - x}{v} \qquad (5\text{-}11)$$

as the phase plane equation.

Equation (5-11) is in the proper form for the application of Lienard's graphical construction,[2] which is perhaps the easiest known technique for treating differential equations of the form of Equation (5-1). The details of the construction are indicated in Figure 5-1.

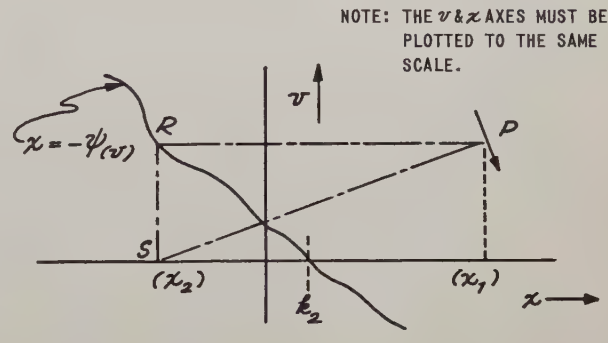


Fig. 5-1. Details of Lienard's construction.

The curve $x = -\psi(v)$ is first plotted. To determine the field direction at any point $P(x,v)$, the procedure is as follows:

From P a line is drawn parallel to the x axis until it cuts the curve $x = -\psi(v)$ at R. From R a perpendicular is dropped to the x axis at S; the field direction at P is then perpendicular to the line SP. The slope of the line SP is $\frac{v}{v + \psi(v)}$, while dv/dx is the negative reciprocal of the slope SP. Note that the sense (as depicted by the arrowhead) is clockwise.

It follows that Lienard's construction can be used to obtain an approximate solution curve in the following way:

1. At the point selected as the initial point, dv/dx is determined graphically, as illustrated.
2. The integral curve in the neighborhood of this point is replaced by a short segment of its tangent.
3. At the end point of this segment the field direction is again determined as outlined in steps 1 and 2.

4. This process is repeated until the complete approximation to the integral curve is obtained. It follows that the approximation can be made quite close by keeping the length of the segments (depicting the field directions) as small as possible.

The details of the construction lead to the following restrictions on the phase plane equation.

1. The coefficient of v in the denominator must be unity.
2. The coefficient of x in the numerator must be -1.
3. The curve $x = -\psi(v)$ * cannot be multivalued on any line parallel to the x axis since the problem arises as to which intercept should be used in the construction.
4. The characteristic phase plane equation is $x = -\psi(v)$. This implies that a characteristic phase plane equation of the form $x = -\psi(x,v)$ is inadmissible. That this is true can be seen by replacing $\psi(v)$ by $\psi(x,v)$ in Figure 5-1. The point P is then dependent on one value of x (for example, x_1) while R is dependent on another value of x (for example x_2). The construction would then yield

$$\frac{dv}{dx_1} = \frac{-\psi(x_2, v) - x_1}{v} \qquad (5\text{-}12)$$

rather than $$\frac{dv}{dx_1} = \frac{-\psi(x_1, v) - x_1}{v} \qquad (5\text{-}13)$$

Hence the construction in invalid.

The realization of good quantitative results with the construction depends only on the length of the tangent segments which approximate the solution curve. (Assuming that the characteristic phase plane curve has been plotted with care.) After a little experience with the construction one soon discovers that the length of the segments depicting the solution curve, at any given point, must become smaller, for any specified accuracy, as the slope of the characteristic phase plane curve increases. The worst case arises when the characteristic phase plane curve has, at any given point, an infinite slope. When this occurs the solution curve is an arc of a circle, which demands that the tangent segments have an infinitesimal length and be of an infinite number. As a practical matter, the difficulty with infinite (or near infinite) slopes can be avoided by swinging arcs of a circle with a compass over regions where the solution is determined by the infinite slopes.

In the following section a technique will be advanced to extend the applicability of Lienard's construction to second-order physical systems which do not have differential equations of the form of Equation (5-1).

* *Henceforth $x = -\psi(v)$ shall be referred to as the characteristic phase plane equation, while the plot of $x = -\psi(v)$ shall be referred to as the chacteristic phase plane curve.*

Section 6

A More Generalized Lienard's Construction

In Section 5, the mechanics of Lienard's construction were discussed and the rather severe restrictions on the co-efficients of x and v in the phase plane equation were set forth. These restrictions required that the differential equation have the form:

$$\ddot{x} + \psi(v) + x = k \qquad (6\text{-}1)$$

This section describes a modified Lienard's construction which permits a more general differential equation than Equation (6-1).

To begin, assume a differential equation of the form:

$$\ddot{x} + b(v)\,\dot{x} + C(x,v)\,x = k \qquad (6\text{-}2)$$

Equation (6-2) yields the following phase plane equation:

$$\frac{dv}{dx} = \frac{-b(v)\,v + k - C(x,v)\,x}{v} \qquad (6\text{-}3)$$

Equation (6-3) is not in the proper form for the application of Lienard's construction. However a slight modification in the regular procedure will provide a simple geometric technique for effecting the solution. Plot, in the phase plane, the following equation:

$$x = -b(v)\,v + k \qquad (6\text{-}4)$$

If the point for which dv/dx is desired is x (for example, point P), then multiply x by $C(x,v)$ to obtain $C(x,v)x$ (for example, point Q). Refer to Figure 6-1.

Referring to Figure 6-1, it is seen that the slope of SQ is:

$$m_{SQ} = \frac{v}{C_{(x,v)}\,x - \left[-b(v)\,v + k\right]} \qquad (6\text{-}5)$$

dv/dx at Q is:

$$\frac{dv}{dx} = \frac{-b_{(v)}\,v + k - C(x,v)\,x}{v} \qquad (6\text{-}6)$$

This slope satisfies Equation (6-3) and can be transferred to P . (Since $ST = QP$, the slope at P can be established directly without performing the construction details at Q .)

For the special case where $C(v,x) \equiv C(x)$, a further simplification results. Notice that $QP = ST = x(1-C(x))$. If a plot of $x(1-C(x))$ *vs* x is constructed to the same scale as the x,v phase plane, then (for any particular value of x) the distance ST can be added or subtracted with a compass. If $x(1-C(x))$ is negative $(C(x)>1)$, add ST toward the left. If $x(1-C(x))$ is positive $(C(x)<1)$, add ST toward the right.

There are no restrictions on the values which

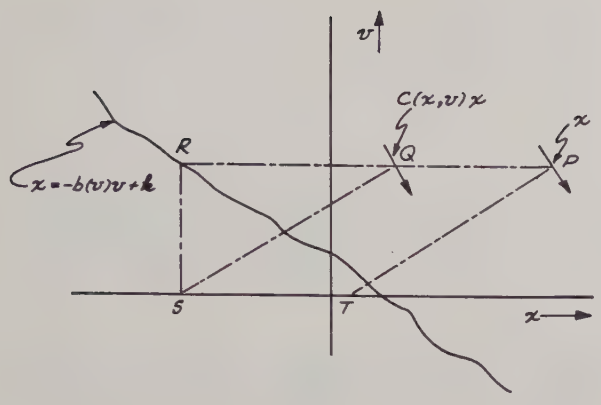


Fig. 6-1. Modified construction details.

the coefficients of Equation (6-6) may assume. (Specifically, the coefficients may even be identically equal to zero. This property is especially useful in analyzing saturation effects in amplifiers and motors.)

The preceding results were exact. As a practical matter, more useful results can be obtained by approximations. Consider the following equations:

$$a(x,v)\ddot{x} + b(x,v)\dot{x} + C(x,v)x = 0 \qquad (6\text{-}7)$$

or

$$\ddot{x} + \frac{b(x,v)}{a(x,v)}\dot{x} + \frac{C(x,v)}{a(x,v)}x = 0 \qquad (6\text{-}8)$$

let

$$\frac{b(x,v)}{a(x,v)} \equiv D(x,v) \qquad (6\text{-}9)$$

and

$$\frac{C(x,v)}{a(x,v)} \equiv E(x,v) \qquad (6\text{-}10)$$

then

$$\ddot{x} + D(x,v)\dot{x} + E(x,v)x = 0 \qquad (6\text{-}11)$$

The phase plane equation is then:

$$\frac{dv}{dx} = \frac{-D(x,v)v - E(x,v)x}{v} \qquad (6\text{-}12)$$

If the variation in $D(x,v)$ due to x is small over a reasonable range of x, then $D(x,v)$ can be evaluated at some convenient point within this range, say x_o (This gives $D(x_o, v)$.

This process can be repeated over as many ranges of x as are needed to obtain a good approximation. Separate characteristic phase plane curves are then plotted in the x,v plane. The construction procedure is the same as before, except that the choice of the characteristic phase plane curve is determined by the particular interval of x in which dv/dx is being determined.

In Section 7, the techniques outlined in this section will be applied to the analysis of a nonlinearity in a high accuracy positional servomechanism. It will serve as an interesting vehicle to clarify some of the mechanics of actually obtaining and plotting the characteristic phase plane equation in the phase plane.

Section 7
Application To A Position Servomechanism

The preceding section described a generalization of Lienard's construction which, as a special case, handles piecewise linear systems with ease. The present section will demonstrate this by applying the technique to the analyses of the performance of an inertially damped position servomechanism with saturation in the forward loop.

The servo under discussion consisted of the following principal components:

1. A high accuracy error transducer (CX-CT pair).
2. A conventional servo amplifier with power output stage operating directly into the motor control-field winding.
3. A size 18 servo motor.
4. A viscous coupled inertial damper.
5. A 40/1 gear train.

The physical configuration is depicted in Figure 7-1, while the system block diagram is given in Figure 7-2.

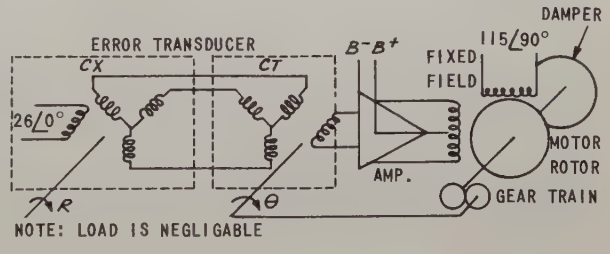


Fig. 7-1. System physical configuration.

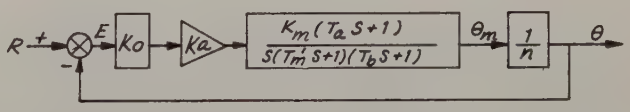


Fig. 7-2. System block diagram.

where: R = input position, rad
θ = output position, rad
E = error, rad

θ_m = motor angular position, rad

K_a = amplifier gain, volt volt^{-1}

K_o = transformation ratio of transducer (CX-CT pair)

K_m = rad volt^{-1} sec^{-1}

T_m' = equivalent motor time constant, sec

T_a & T_b are time constants introduced by the inertial damper, sec

n = gear ratio between output shaft and motor shaft

S = LaPlace operator

In this paper the burden of deriving Figure 7-2 will not be shouldered (Reference 1 contains an excellent derivation). It will be assumed that the system has been synthesized to satisfy some linear criterion, with the values depicted in Figure 7-3 resulting.

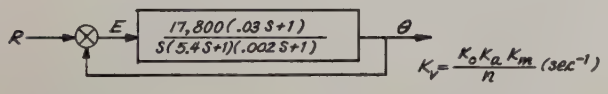


Fig. 7-3. Final system block diagram.

The Bode plot for this system is given in Figure 7-4.

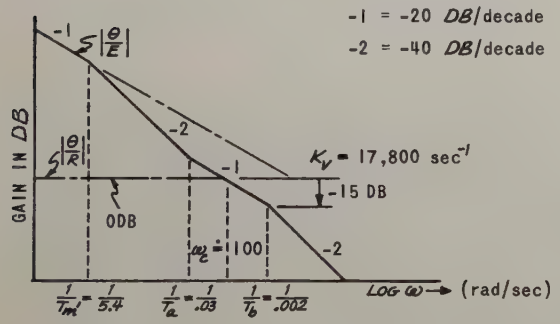


Fig. 7-4. System Bode plot.

It will be necessary to approximate the third order block diagram of Figure 7-3 with a second order one, since the non-linear technique advanced in Section 6 is applicable only to second order systems. Inspection of Figure 7-4 makes an excellent approximation obvious. The pole at $S = -1/T_b$ corresponds to a gain of approximately -15 DB, hence it can be ignored.* This approximation gives the block diagram shown in Figure 7-5.

Consider the case of saturation. In this particular servo, saturation in the amplifier occurs before torque saturation in the servo motor. Because

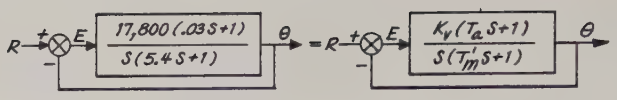


Fig. 7-5. Approximate block diagram.

of this, the amplifier can no longer be represented as a constant. Replace the constant by a non-linear function $(A(e))$ in the block diagram, giving Figure 7-6.

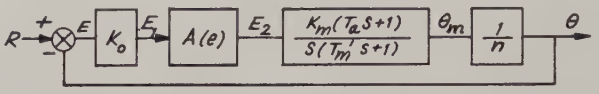


Fig. 7-6. Block diagram with nonlinear element.

The non-linear gain function $(A(e))$ can be plotted versus any variable desired** Hence it is possible to construct plots of $A(e)$ vs. θ, R or E. A plot of $A(e)$ vs. E will be selected since it is desirable in servo work to use an error rate versus error phase plane. This particular phase plane is convenient because a zero steady state error corresponds to the origin of the phase plane.

If the idealized and normalized plot of E_2 vs. E given in Figure 7-7 is taken as the transfer function from E_2 to E, then the block diagram of Figure 7-8 can be used.

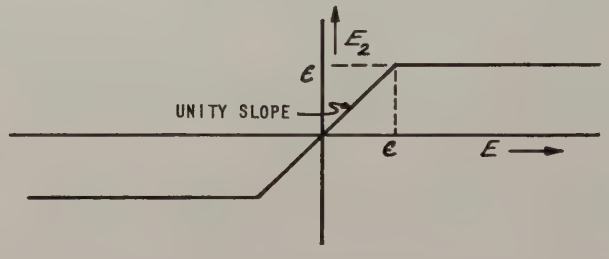


Fig. 7-7. Idealized saturation curve.

Note: This particular servo amplifier saturates for an arc input of eight minutes.

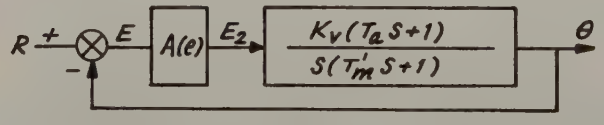


Fig. 7-8. Normalized block diagram.

* *The effect of the pole at $S = -1/T_b$ can be investigated by applying the more general discussion of Appendix A.*

** *This is not always true, but here we have a unity feedback system.*

When

$$-\epsilon < E < \epsilon \quad , \quad A(e) = 1$$

$$\frac{\theta}{E} = \frac{K_v(T_a S+1)}{S(T_m' S+1)} \tag{7-1}$$

$$\therefore \frac{E}{R} = \frac{1}{1+\theta/E} = \frac{T_m' S^2+S}{T_m' S^2+(1+K_v T_a)S+K_v} \tag{7-2}$$

or

$$T_m' S^2 E + (1+K_v T_a) SE + K_v E = R\left[T_m' S^2 + S\right] \tag{7-3}$$

Let $S \equiv d/dt$ and limit R to k (a constant).

$$T_m' \frac{d^2e}{dt^2} + (1+K_v T_a)\frac{de}{dt} + K_v e = 0 \tag{7-4}$$

It is convenient to work with ratios of T_m'/K_v, since this ratio is essentially a constant.

$$\frac{T_m'}{K_v}\frac{d^2e}{dt^2} + \left(\frac{1+K_v T_a}{K_v}\right)\frac{de}{dt} + \frac{K_v}{K_v} e = 0 \tag{7-5}$$

$$\text{if } K_v T_a >> 1 \tag{7-6}$$

$$\frac{T_m'}{K_v}\frac{d^2e}{dt^2} + T_a\frac{de}{dt} + e = 0 \tag{7-7}$$

$$\text{let } t_1 = \sqrt{K_v/T_m'}\, t \,, \therefore \frac{de}{dt} = \sqrt{K_v/T_m'}\frac{de}{dt_1}$$

$$\text{or } \dot{e} = \sqrt{K_v/T_m'}\,\dot{e}_1 \tag{7-8}$$

This leads to Equation (7-9)

$$\frac{d^2e}{dt_1^2} + T_a\sqrt{K_v/T_m'}\frac{de}{dt_1} + e = 0 \tag{7-9}$$

The phase plane equation is (letting $de/dt_1 = \dot{e}_1$):

$$\frac{d\dot{e}_1}{de} = \frac{-T_a\sqrt{K_v/T_m'}\,\dot{e}_1 - e}{\dot{e}_1} \tag{7-10}$$

Substitution of the numerical values given in Figure 7-5 gives:

$$\frac{d\dot{e}_1'}{de} = \frac{-1.72\,\dot{e}_1 - e}{\dot{e}_1} \quad \text{when} \quad -\epsilon < e < \epsilon \tag{7-11}$$

When $E > +\epsilon$ (or $E < -\epsilon$) E_2 in Figure 7-8 is a constant and equal to $\pm\epsilon$ (depending on the sign of E).

$$\therefore \frac{\theta}{\epsilon} = \frac{K_v(T_a S+1)}{S(T_m' S+1)} \tag{7-12}$$

and $R - \theta = E$

$$R - \pm\epsilon\frac{K_v(T_a S+1)}{S(T_m' S+1)} = E \tag{7-13}$$

or

$$R\left[T_m' S^2 + S\right] - \pm\epsilon\left[K_v(T_a S+1)\right] = T_m' S^2 E + SE \tag{7-14}$$

Let $S \rightarrow \frac{d}{dt}$, $R = k$

$$T_m'\frac{d^2e}{dt^2} + \frac{de}{dt} \pm \epsilon K_v = 0 \tag{7-15}$$

To conform to the time scale used in Equation (7-11), divide Equation (7-15) by K_v,

$$\frac{T_m'}{K_v}\frac{d^2e}{dt^2} + \frac{1}{K_v}\frac{de}{dt} = \mp\epsilon \tag{7-16}$$

It would appear that the generalized Lienard's construction cannot be applied to Equation (7-16) since the error term is missing. However, as pointed out in Section 6, the non-linear coefficients may be explicitly equal to zero, which allows for the replacement of Equation (7-16) with Equation (7-17).

$$\frac{T_m'}{K_v}\frac{d^2e}{dt^2} + \frac{1}{K_v}\frac{de}{dt} + ge = \mp\epsilon \tag{7-17}$$

where $g \equiv 0$

This gives the following phase plane equation:

$$\frac{d\dot{e}_1}{de} = \frac{-\frac{1}{K_v}\sqrt{K_v/T_m'}\,\dot{e}_1 - ge \mp \epsilon}{\dot{e}_1}$$

$$\text{for } e > \epsilon \,,\, e < -\epsilon \qquad \text{where } g \equiv 0 \tag{7-18}$$

Substitution of the numerical values shown in Figure 7-5 gives:

$$\frac{d\dot{e}_1}{de} = \frac{-.0034\,\dot{e}_1 - ge \mp \epsilon}{\dot{e}_1} \quad g \equiv 0 \tag{7-19}$$

Equations (7-11) and (7-19) will completely specify the systems performance in an $\dot{e}_1$ vs. e phase plane once the initial values have been determined. Notice that initial values are of interest here, since our basic equations were derived from the block diagram of Figure 7-8, which automatically makes all initial conditions zero. To determine the initial values when $-\epsilon \le k \le \epsilon$, rewrite Equation (7-2) as

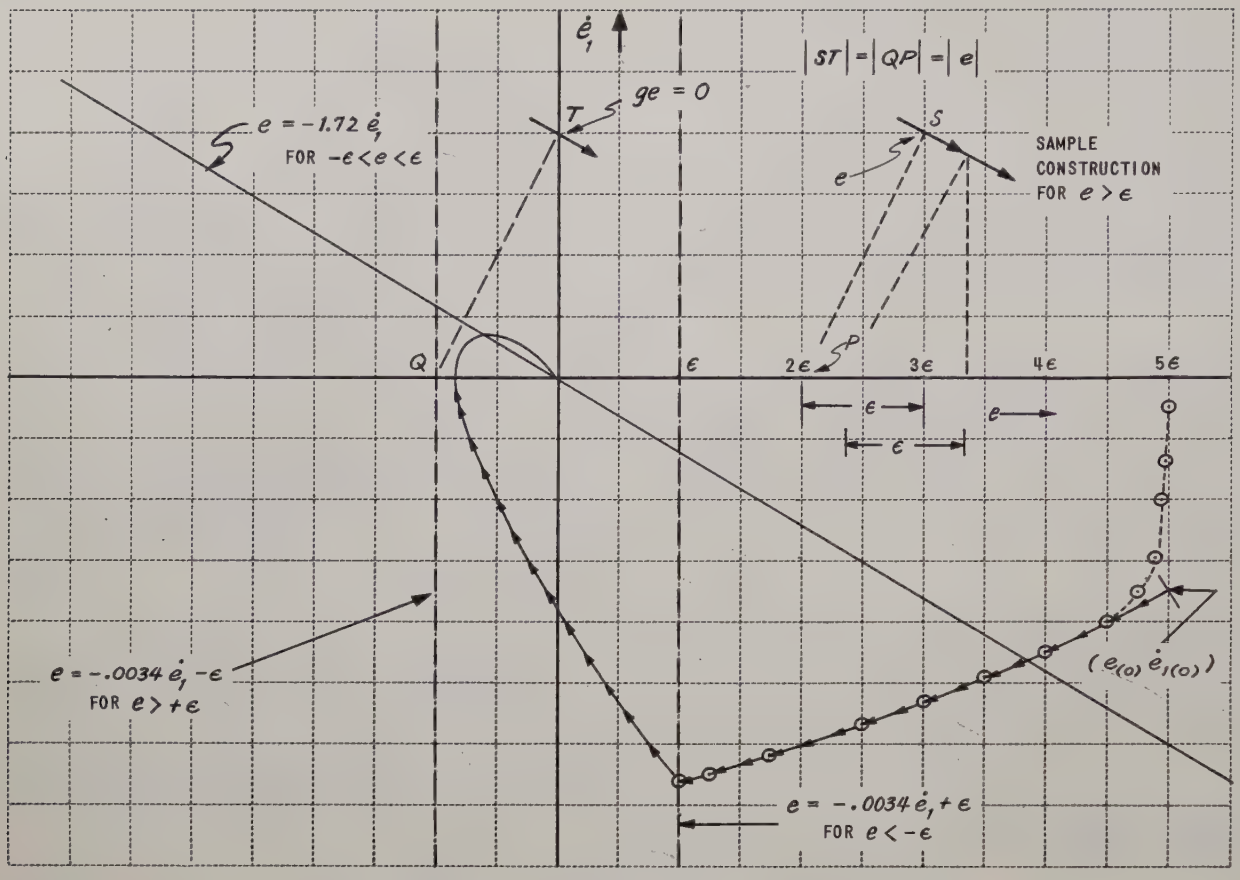


Fig. 7-9. Saturation construction.

$$E(S) = \frac{k\,(S + 1/T_m')}{\left[S^2 + \left(\frac{1 + K_V T_a}{T_m'}\right)S + \frac{K_V}{T_m'}\right]} \qquad (7\text{-}20)$$

with R being replaced by k/S. Equation (7-20) can be written in the following form:*

$$E_{(S)} = k\left[\frac{1}{S} - \frac{K_V T_a}{T_m' S^2} - \left(\frac{K_V}{T_m} - \frac{K_V T_a}{T_m'^2} - \frac{K_V^2 T_a^2}{T_m'^2}\right)\frac{1}{S^3} - \cdots\right] \qquad (7\text{-}21)$$

or

$$e_{(t)} = k - \frac{k K_V T_a}{T_m'}t - k\left(\frac{K_V}{T_m'} - \frac{K_V T_a}{T_m'^2} - \frac{K_V^2 T_a^2}{T_m'^2}\right)\frac{t^2}{2} - \cdots \qquad (7\text{-}22)$$

Evaluating Equation (7-22) at $t = 0$ gives the initial value of e :

$$e_{(o)} = k \qquad (7\text{-}23)$$

Differentiating Equation (7-22) and evaluating the result at $t = 0$ gives the initial value of de/dt as:

$$\dot{e}_{(o)} = -\frac{k K_V T_a}{T_m'} \qquad (7\text{-}24)$$

But

$$\frac{de}{dt_1} = \sqrt{T_m'/K_V}\ \frac{de}{dt}\ , \qquad (7\text{-}25)$$

therefore

$$\dot{e}_{1(o)} = -k T_a \sqrt{K_V/T_m'} \qquad (7\text{-}26)$$

When $k > \epsilon$ (or $< -\epsilon$), the same sort of analysis applied to Equation (7-13) yields

$$e_{(o)} = k \qquad (7\text{-}27)$$

and

$$\dot{e}_{1(o)} = \mp\, \epsilon T_a \sqrt{K_V/T_m'} \qquad (7\text{-}28)$$

The use of Equations (7-10) and (7-18) are

* *The form of Equation (7-20) precludes the use of the initial value theorem in determining the initial value of de/dt .*

demonstrated in Figure 7-9 for a 5ϵ step input. Since $\dot{e} > +\epsilon$, Equations (7-27) and (7-28) determine the initial values. The tangent segments were made relatively large for the sake of clarity in following the construction details. In addition, a sample construction was carried out, in the upper half plane, to demonstrate the construction details for $e > \epsilon$.

Notice that the construction becomes trivial for $e > \epsilon$ (or $< -\epsilon$), due to the steepness of the characteristic phase plane curves. This can be seen more clearly by writing Equation (7-19) in an approximate form:

$$\frac{d\dot{e}_1}{de} = \frac{\mp \epsilon}{\dot{e}_1} \qquad (7\text{-}29)$$

An inspection of Equation (7-29) makes it clear that the basic slope idea advanced in Section 1 works very nicely; it is only necessary to measure off ϵ , in the proper direction, on the e axis to construct the tangential approximation. (Refer to Figure 7-9)

More generally, with $g = 0$, the discussion of Section 4 can be applied directly to Equation (7-19), since the numerator is a function of $\dot{e}_1$ only.

The circled points on Figure 7-9 denote part of the solution that would have resulted if the pole at $S = -\frac{1}{T_6}$ had not been ignored.

Bibliography

J. JURSIK, J.F. KAISER, and J.E. WARD, "A Dual-Made Damper-Stabilized Servo", Transactions of the ASME, Paper No. 56 - IRD-6; 1956.

J.J. STOKER, "Nonlinear Vibrations", Interscience Publishers, Inc.; 1950 (Pages 31 and 32 for Lienard's graphical construction).

Appendix A

A Third Order Differential Equation

This appendix will consider a third order equation, the object being to illustrate the generality of the ideas used in Section 1. Consider the general third order differential equation

$$A\frac{d^3x}{dt^3} + B\frac{d^2x}{dt^2} + C\frac{dx}{dt} + Dx = f(t) \qquad (A\text{-}1)$$

where A , B , C and D are coefficients dependent on a , v , x and t .

$f(t)$ = forcing function

and a = acceleration

v = velocity

x = displacement

t = time

Equation (A-1) can be re-written as follows:

$$\frac{d^3x}{dt^3} = \frac{f(t)}{A} - \frac{B}{A}\frac{d^2x}{dt^2} - \frac{C}{A}\frac{dx}{dt} - \frac{D}{A}x \qquad (A\text{-}2)$$

or $$\frac{da}{dt} = \frac{f(t)}{A} - \frac{B}{A}a - \frac{C}{A}v - \frac{D}{A}x \qquad (A\text{-}3)$$

For the convenience of the discussion, let the right hand side of Equation (A-3) be called $-g(a, v, x, t)$. Then

$$\frac{da}{dt} = -g(a, v, x, t) \qquad (A\text{-}4)$$

However, we have additional information at our disposal in the form of defining equations:

$$\frac{dx^3}{dt^3} = \frac{da}{dt} = a\frac{da}{dv} = v\frac{da}{dx} \qquad (A\text{-}5)$$

and $$\frac{dv}{dt} = a \quad , \quad \frac{dx}{dt} = v \qquad (A\text{-}6)$$

Using Equations (A-5) and (A-6) with Equation (A-4) gives equations which can be interpreted as slopes in their various planes (for example, Equation (A-11) defines the slope of a solution curve in the acceleration, velocity plane).

$$\frac{dv}{dt} = a \qquad (A\text{-}7)$$

$$\frac{dx}{dt} = v \qquad (A\text{-}8)$$

$$\frac{dv}{dx} = \frac{a}{v} \qquad (A\text{-}9)$$

$$\frac{da}{dt} = -g(a, v, x, t) \qquad (A\text{-}10)$$

$$\frac{da}{dv} = -g(a, v, x, t)/a \qquad (A\text{-}11)$$

$$\frac{da}{dx} = -g(a, v, x, t)/v \qquad (A\text{-}12)$$

Equations (A-7) through (A-12) can be used, as Equations (1-4) through (1-6) were used in Section 1, to obtain approximations to the solution curves in the appropriate plane. Figure A-1 demonstrates a possible way to set up the various planes to make maximum use of parallel lines between planes to construct slopes (note that only three of the planes are required to construct a complete solution).

A sample construction has been included in Figure A-1 to demonstrate the simplicity of the scheme. Notice that once $g(a_0, v_0, x_0, t_0)$ has been computed that all the tangential approximations can be easily constructed with parallel lines and/or a compass.

It is obvious that even higher order equations can be handled in much the same manner; of course there will be a comensurate increase in the number of planes required as the order of the equation increases.

As in Section 1, constructions in seemingly redundant planes will serve to enhance the accuracy of the technique.

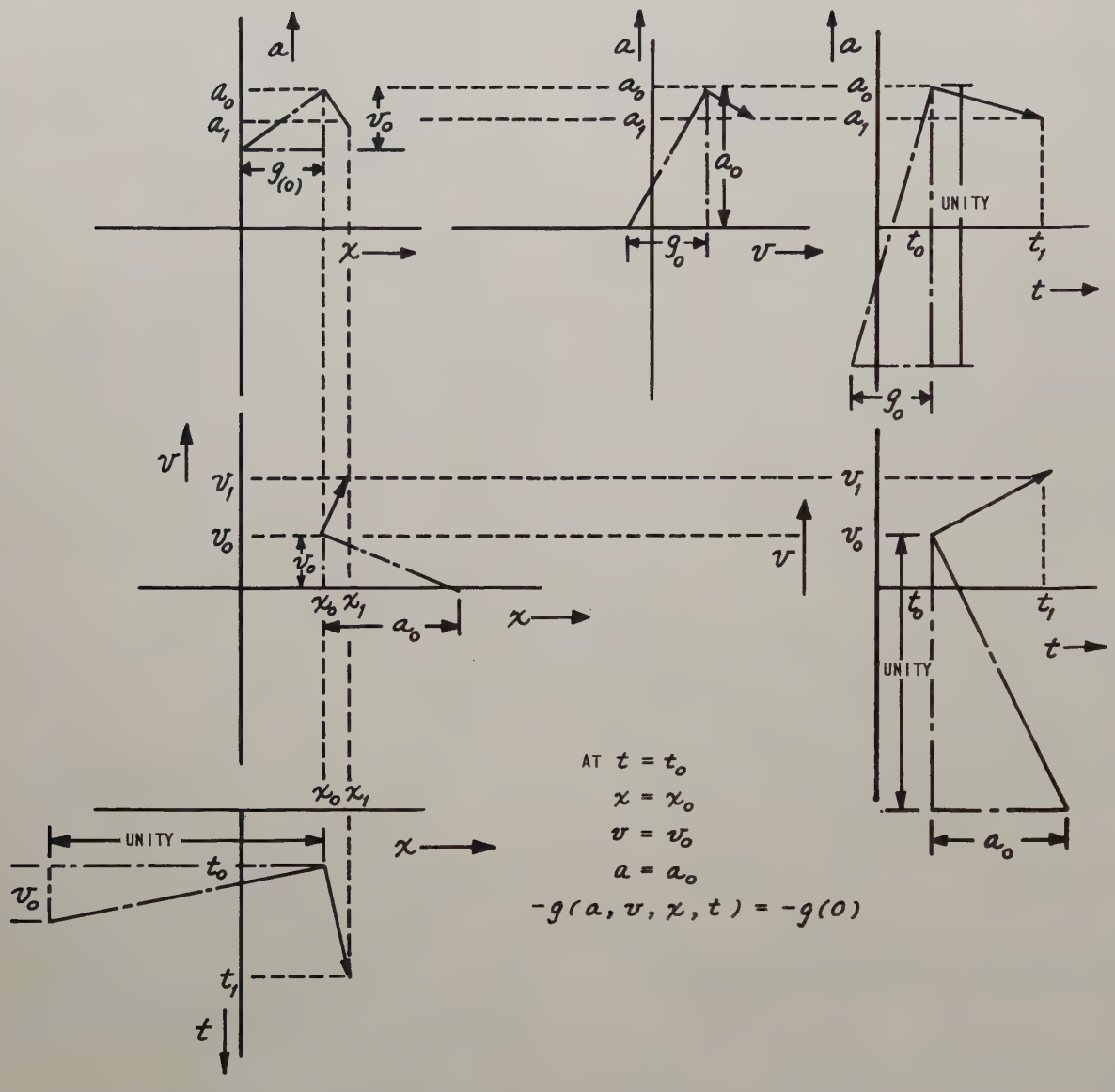


Fig. A-1. Construction details for a third order system.

ON THE USE OF GROWING HARMONIC EXPONENTIALS TO IDENTIFY STATIC NONLINEAR OPERATORS

H. J. Lory, D. C. Lai, W. H. Huggins
Department of Electrical Engineering
The Johns Hopkins University
Baltimore 18, Maryland

Summary

The following paper describes a method of obtaining a polynomial characteristic function for a nonlinear static system. This function, $F(x) = hx + mx^2 + dx^3$, is obtained by the application of a growing exponential $x = \exp(t)$ to the input of the system and the filtering of the output $h\exp(t) + m\exp(2t) + d\exp(3t)$, into its separate components $h\exp(t)$, $m\exp(2t)$, and $d\exp(3t)$. The values of these three components at $t = 0$ are the polynomial coefficients h, m, and d respectively. The identification of systems not exactly describable by a cubic gives rise to an error minimization problem; the technique described in this paper minimizes the weighted mean-square error, with a weighting function $1/x$. This method is compared with the more widely known sinusoidal analysis of nonlinear systems. Experimental results are given.

Introduction

There has been a growing tendency in recent years to regard a given system not as a set of physical components, but as a collection of mathematical operators acting on the input signals to give the outputs. This approach, while tending to mask much detailed understanding of the internal behavior of the system, sharpens insight into the transformation of the input signal. Also, there is an opportunity for considerable saving in the time and effort required to analyze the system.

To attain this economy, various classes of operators suitable for representing systems must be established. Even more important to the engineer, experimental methods must be devised for practical measurement of the parameters that characterize physical operators.

This paper discusses an experimental method for measuring the parameters of a polynomial representation of a static nonlinear operator. The method uses a growing exponential input signal as a testing function. Because all growing exponential signals are eigenfunctions of all stable, linear, stationary systems, there is a possibility that this method of measuring system nonlinearities may be extended to dynamic systems as well. However, the present paper considers only static systems in which the output at any instant is a single-valued function of the input at the same instant.

In this paper, it will be assumed that the system output may be approximated by a polynomial function of the input. The degree of the polynomial will generally depend upon the nature of the nonlinear characteristic and upon the accuracy with which it is to be represented. In many instances, a cubic will provide adequate accuracy, and it is this case that we consider here. That is, for an input x to the system, the output can be represented with sufficient accuracy by $hx + mx^2 + dx^3$. (By proper choice of the output zero level, the response for $x = 0$ may be made to vanish). The method is, in principle, applicable to quartic and higher-degree representations. However, the precision required of the measuring instrumentation may limit the method to polynomials of fairly low degree.

An Ideal Filter for Exponential Components

Zadeh has pointed out[1] that the basic notion of signal filtration is a much more general operation than most engineers realize, and that many of our common notions concerning the separation of a complex wave into its various frequency components are applicable to a wider class of functions than is generally considered. In particular, since the eigenfunctions of all stationary, stable, linear systems include the class of growing exponentials in addition to the widely used sinusoid, it is possible to construct an "ideal" filter for decomposing a signal into a sum of a discrete set of growing exponentials.

The ideal filter shown in Figure 1 for the set of growing harmonic exponential components $\exp(t)$, $\exp(2t)$, and $\exp(3t)$ has already been described by one of the authors.[2]

If an exponential of the form $\exp(t)$ is applied at $t = -\infty$ to the input of a nonlinear system characterized by a cubic, the output will be $h\exp(t) + m\exp(2t) + d\exp(3t)$ (Figure 2). If this output signal is now applied to the input of the ideal filter, $h\exp(t)$ will appear at the first port, $m\exp(2t)$ at the second, and $d\exp(3t)$ at the third, as shown in Figure 2. Furthermore, the three outputs of the ideal filter will provide separately the three polynomial coefficients, for if each is sampled when the value of the input, $\exp(t)$, to the nonlinear system is unity (corresponding to the instant $t = 0$), the values of the three outputs at that same instant will be $h\exp(0)$, $m\exp(2\cdot0)$, and $d\exp(3\cdot0)$; or h, m, and d, respectively. Thus, by applying an exponential $\exp(t)$ to the input of a nonlinear system, passing the system output

through an ideal exponential filter, and sampling the outputs of the filter at the proper instant, one obtains a direct measurement of the coefficients in the characteristic function that describes the nonlinear system.

Approximation of Arbitrary Nonlinear Characteristics

In the foregoing discussion, it was assumed that the system under consideration had a characteristic function of the form $hx + mx^2 + dx^3$. What if the characteristic function is not of this specialized form? Then, we might still obtain a useful representation by approximating the actual nonlinear characteristic with a cubic function, using some criterion for selecting the "best" approximation. Least-squares techniques may be applied to numerical data points along the input-output characteristic to fit a smooth curve to these points. This generally requires the observation of many more data points than there are parameters to be determined. However, it is also possible to use simple signal filtration methods to measure directly the parameters which describe the "best fitting" curve. In particular, it is possible to determine the coefficients of the best-fitting cubic approximation to an arbitrary nonlinear system characteristic by observing the response of that system to a single growing exponential exp(t).

When the system nonlinearity is exactly representable by a simple cubic, the response of the system to a single growing exponential input will consist exactly of the sum of three growing exponential terms. As indicated in Figure 2, the ideal filter separates these three components so that the output signals at the three terminals are h exp(t), m exp (2t), and d exp (3t) respectively, where it will be recalled that h, m, and d are constants in time which characterize the nonlinear system.

If, however, the output of the nonlinear system is given by an arbitrary F(x) when the input has the value x, the response of the system to an exponential input $x = \exp(t)$ will be $F[\exp(t)] = f(t)$, where f(t) can usually no longer be represented exactly as the sum of a finite number of exponential components. However, f(t) can still be approximated within some error by the same three exponential terms previously used, viz.,

$$f(t) \simeq h' \exp(t) + m' \exp(2t) + d' \exp(3t) = f'(t), \quad (1)$$

where $f'(t)$ is the approximate expression obtained for specified values of the coefficients h', m', and d'. It is clear that these coefficients should be chosen so that the discrepancy between f(t) and $f'(t)$ is small. The choice of the h', m', and d' parameters depends upon the particular error criterion that is used. We have used a uniform weighting of the error E over time:

$$E(t) = \int_{-\infty}^{t} \left[f(\tau) - f'(\tau) \right]^2 d\tau , \quad (2)$$

and have designed our instrumentation so as to yield at any time, t, coefficients that will minimize the error in a least-square sense over the interval $-\infty < \tau < t$.

The foregoing discussion has been concerned with the curve-fitting of a time function. However, our primary interest is to reconstruct a cubic in x which is an approximation to the output F(x) in terms of the input x.

Let $x = e^t$. Then, $x^2 = e^{2t}$, $x^3 = e^{3t}$, and

$$dx = e^t dt = x\, dt, \text{ or } dt = \frac{1}{x} dx \quad (3)$$

$$f(t) = F(e^t) = F(x) . \quad (4)$$

The error integral when expressed as a function of x becomes,

$$E(o) = \int_0^{'} \left[F(x) - F'(x) \right]^2 dx/x = \int_0^{'} \frac{\left[F(x) - h'x - m'x^2 - d'x^3 \right]^2}{x} dx . \quad (5)$$

Minimization of the mean-square error in approximating f(t) thus leads to a polynomial least-square approximation in x with a weighting factor 1/x. It might seem that a 1/x weighting factor renders the approximation inappropriate. However, this is not necessarily so. One may frequently require that the relative accuracy of the approximation be maintained for signal amplitudes that are considerably smaller than the maximum amplitude used in determining the approximation. The 1/x weighting factor achieves an approximation that has this desirable property.

Realization of Instrumentation

The most critical part of the instrumentation is the ideal filter, since the problem of generating exponentials and of simultaneous sampling of several analogue voltages is readily solved by well known methods.

Consider now the device shown in Figure 3. It may be constructed using ordinary analogue computer elements.[2,4] If a unit impulse is applied to the input terminal, the respective signals at the three output terminals for $t > 0$ are:

$$\varphi_1 = \sqrt{2}\, e^{-t} ,$$
$$\varphi_2 = \sqrt{4}\, (-2e^{-t} + 3e^{-2t}), \quad (6)$$
$$\varphi_3 = \sqrt{6}\, (3e^{-t} - 12e^{-2t} + 10\, e^{-3t}) .$$

These signals obviously are identically zero for $t < 0$.

It may be shown[2] that these signals are orthonormal over the interval from 0 to ∞. Knowing the impulse response of the filter, one may use convolution to calculate the output $c_k(t)$ of the output port for any input f(t).

$$c_k(t) = \int_0^{\infty} f(t - \tau)\, \varphi_k(\tau) d\tau . \quad (7)$$

Let $t' = -\tau$. Then (7) becomes

$$c_k(t) = \int_{-\infty}^{0} f(t + t')\,\varphi_k(-t')\,dt' . \qquad (8)$$

Now, consider the time-reversal signals

$$\tilde{\varphi}_k(t) = \varphi_k(-t) . \qquad (9)$$

These functions $\tilde{\varphi}_k$ are orthonormal over the interval $-\infty \to 0$; unlike the original set, they consist of growing rather than decaying exponentials. The coefficients of the growing exponentials, however, are the same. Let us substitute (9) into (8) and evaluate at $t = 0$; we obtain

$$c_k(0) = \int_{-\infty}^{0} f(t')\,\tilde{\varphi}_k(t')\,dt' . \qquad (10)$$

This, however, is the generalized Fourier coefficient of the decomposition of the signal $f(t)$ in the approximation [5]

$$f(t) = c_1 \tilde{\varphi}_1(t) + c_2 \tilde{\varphi}_2(t) + c_3 \tilde{\varphi}_3(t). \qquad (11)$$

In other words, by sampling the output of the above filter at the proper epoch, normally at $t = 0$, the coefficients c_1, c_2, and c_3 of the approximation of the input signal may be measured directly. The orthonormal set $\{\tilde{\varphi}_k(t)\}$ is composed of the same harmonic exponentials used in our previous discussion. Moreover, because of the orthonormality of these functions, the approximation minimizes the mean-square error. But this approximating function may also be expressed as a linear combination of $\exp(t)$, $\exp(2t)$, and $\exp(3t)$. Hence, the right-hand side of (11) can be written in the form

$$he^{t} + me^{2t} + de^{3t} = c_1 \tilde{\varphi}_1 + c_2 \tilde{\varphi}_2 + c_3 \tilde{\varphi}_3, \qquad (12)$$

where

$$\begin{aligned} \tilde{\varphi}_1 &= \sqrt{2}\, e^{t} , \\ \tilde{\varphi}_2 &= \sqrt{4}(-2e^{t} + 3e^{2t}), \\ \tilde{\varphi}_3 &= \sqrt{6}(3e^{t} - 12e^{2t} + 10e^{3t}) . \end{aligned} \qquad (13)$$

By substituting (13) into (12) and equating coefficients, one obtains h, m, and d in terms of c_1, c_2, and c_3,

$$\begin{aligned} h &= \sqrt{2}\, c_1 - 4c_2 + 3\sqrt{6}\, c_3 , \\ m &= 6c_2 - 12\sqrt{6}\, c_3 , \\ d &= 10\sqrt{6}\, c_3 . \end{aligned} \qquad (14)$$

Simple weighting circuits may be designed to combine the c_1, c_2, and c_3 voltages so as to obtain signals that when sampled at $t = 0$ will yield the values of the h, m, and d coefficients (see Figure 4). Moreover, the outputs R_1, R_2, and R_3 will, at any instant t, be the present values of the exponential components of the approximation which is "best" in a least-squares sense over all past time. The arrangement of Figure 4 is thus an ideal filter for the growing exponentials of Equation (12).

Experimental Results

To generate the set of growing harmonic exponentials, an operational amplifier having the transmittance, $\frac{1}{s-1}$, was used to produce a growing exponential e^{t}. The harmonic exponentials $\exp(2t)$ and $\exp(3t)$ were produced by "squaring" and "cubing" $\exp(t)$ using analogue function multipliers. Unfortunately, these function multipliers were imperfect and the result of "squaring" $\exp(t)$ did not yield $\exp(2t)$ exactly. Similarly, the "cubing" of $\exp(t)$ did not yield $\exp(3t)$ exactly. The ideal filter previously described may be used to determine the error of these "squaring" and "cubing" operations, as will now be shown.

Figure 6 shows the result x_2 of "squaring" x. Because the "squaring" operation is not exact, the actual nonlinearity may be approximated by a large square-law term plus small linear and cubic error terms. The values of these error terms will depend upon the maximum value x_0 of the range $0 < x < x_0$ over which the approximation is made. For instance, taking $x_0 = 1$, we may read the ordinates from the data of Figure 6 to find that

$$x_2 = 0.03x + 1.00x^2 - 0.01x^3 . \qquad (15)$$

Likewise, if the output x_3 of the "cubing" device is analyzed, yielding the data of Figure 7, it is found that linear and quadratic terms are present. Over the interval $0 < x < 1$, the best fitting approximation of the "cubing" characteristic is found to be

$$x_3 = -0.015x + 0.135x^2 + 0.88x^3, \qquad (16)$$

where the coefficients are given by the values of R_1, R_2 and R_3 at $x = 1$.

We may now create an arbitrary non-linear device by combining with x the outputs x_2 and x_3 of the "squaring" and "cubing" circuits after multiplication by constant scale factors. In this way, it is a simple matter to synthesize the non-linear characteristic

$$F_A(x) = 0.2x_1 - 0.6x_2 + 0.4x_3. \qquad (17)$$

This has been plotted at the top of Figure 8(A). Also shown in Figure 8(A) is the result of analyzing $F_A(\exp t)$. The best-fitting approximation of $F_A(x)$ over the interval $0 < x < 1$ is found by measuring the values of R_1, R_2 and R_3 at $x = 1$. This gives

$$F_A(x) = 0.17x - 0.54x^2 + 0.38x^3. \qquad (18)$$

The fact that the experimentally determined Equation (18) does not agree with Equation (17)

is explained, in part, by the fact that the squaring and cubing devices were imperfect. If the more accurate expressions for x_2 and x_3, as given by Equations (15) and (16), are substituted into Equation (17), we find that the synthesized nonlinearity should be expressed by

$$\begin{aligned} 0.2x_1 &= 0.200x \\ -0.6x_2 &= -0.018x - 0.600x^2 + 0.006x^3 \\ +0.4x_3 &= -0.006x + 0.054x^2 + 0.352x^3 \\ \hline F_A(x) &= 0.176x - 0.546x^2 + 0.358x^3 \quad . \end{aligned} \tag{19}$$

Thus, the discrepancy between (17) and (18) may be accounted for by the imperfections in the squaring and cubing circuits used.

In Figure 8(B), the non-linear characteristic has been modified simply by changing the algebraic signs of the x_2 and x_3 components, viz.

$$F_B(x) = 0.2x_1 + 0.6x_2 - 0.4x_3 \tag{20}$$

Also shown in Figure 8(B) is the result of analyzing $F_B(\exp t)$. The best fitting approximation over the interval $0 < x < 1$ is found, as before, to be

$$F_B(x) = 0.22x + 0.57x^2 - 0.38x^3 \quad . \tag{21}$$

Here also, the discrepancy between Equations (8)(20) and (21), may be largely explained by the imperfections of the squaring and cubing devices. Substitution of Equations (15) and (16) into (20) yields

$$\begin{aligned} 0.2x_1 &= 0.200\,x \\ 0.6x_2 &= 0.018x + 0.600x^2 - 0.006x^3 \\ -0.4x_3 &= 0.006x - 0.054x^2 - 0.352x^3 \\ \hline F_B(x) &= 0.224x + 0.546x^2 - 0.358x^3 \quad . \end{aligned} \tag{22}$$

Next, consider a nonlinearity, such as the two line segments shown in Figure 9(A), which is not exactly describable by a third degree polynomial. The three component exponentials are shown in Figure 9(B). Note that since the nonlinear characteristic is linear for inputs between $x = 0$ and 0.5, only the term $x = \exp(t)$ is present in this region. For $x \geq 0.5$, however, the other two components are needed to achieve a good approximation of the nonlinearity. By sampling the values of R_1, R_2, and R_3 at $x = 1$, the values of h, m, and d are obtained, yielding the polynomial approximation.

$$F(x) = 0.456x_1 - 0.086x_2 + 0.694x_3. \tag{23}$$

The approximation given by (23) was generated experimentally and is shown in Figure 10, together with the original characteristic function. Although the approximation is fairly close throughout the interval $0 < x \leq 1$, it is especially good near the origin because of the weighting factor $\frac{1}{x}$ in the error integral.

Comparison with Sinusoidal Distortion Measurements

In this section, we consider the relationship between the measurement of a nonlinear characteristic by using growing exponentials, and the more familiar harmonic distortion that would be produced by this nonlinearity if the input were a sinusoid of some specified amplitude, A.

When a sinusoid is applied to a nonlinear system, the harmonic components produced at the output may be measured individually by selective band-pass filters. From these data, the characteristic function associated with the nonlinear system may be determined.[6]

There are several advantages of using a growing exponential as a testing function instead of a sinusoidal signal. First, the outputs from the ideal filter may be sampled at any instant, say at $t = t_0$, to obtain an approximation over the interval from $x = 0$ to $x = x_0 = e^{t_0}$. In a single experiment, one may record the quantities R_1, R_2 and R_3 as continuous functions of t_0, thus obtaining an infinite number of best approximations. To obtain a similar set of data using sinusoids, one has to repeat the experiments at a variety of input amplitudes. Second, to achieve a close fit near the origin, the exponentials are much better suited than sinusoids because the exponential approximation involves a weighting factor $\frac{1}{x}$, which weights the errors at large amplitudes less heavily than it does at small amplitudes; whereas the sinusoidal approximation involves a weighting factor $\frac{1}{\sqrt{1 - x^2}}$ which weights heavily the errors near the peak amplitude. Third, for representing rectifying systems such as diodes, the exponentials are much better suited than sinusoids. Since the sinusoid will cause the input to vary over the range $-A \leq x \leq A$, we must first specify the nonlinearity for both positive and negative values of the input. On the other hand, a growing exponential signal has only one polarity, and it is therefore necessary to make an additional measurement using a growing exponential of opposite sign to identify the nonlinear characteristic. It may then be expressed over both ranges as

$$F(x) = \begin{cases} h^+x + m^+x^2 + d^+x^3, & 0 \leq x \leq 1 \\ h^-x + m^-x^2 + d^-x^3, & -1 \leq x \leq 0, \end{cases} \tag{24}$$

where as before, we have chosen a unit interval on either side of the origin over which to optimize the cubic approximation. Now the Equation (24) may be used to calculate the harmonic distortion that would occur if an input $x = \sin \omega t$ were transformed by the nonlinearity $F(x)$. This establishes a link between the distortion coefficients h, m, and d as measured by our method, and the more familiar measure of harmonic distortion. Because of the quarter-wave symmetry of the output of a static nonlinearity, the output will have the specialized form:

$$a_o + a_1 \sin\omega t + a_3 \sin 3\omega t + b_2 \cos 2\omega t \ . \quad (24)$$

A linear relation between these various coefficients is given by the matrix equation

$$\begin{bmatrix} a_o \\ a_1 \\ a_3 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0.318 & 0.250 & 0.212 & -0.318 & 0.250 & -0.212 \\ 0.500 & 0.424 & 0.375 & 0.500 & -0.424 & 0.375 \\ 0.000 & -0.042 & -0.125 & 0.000 & +0.042 & -0.125 \\ -0.212 & -0.250 & -0.255 & 0.212 & -0.250 & 0.255 \end{bmatrix} \begin{bmatrix} h^+ \\ m^+ \\ d^+ \\ h^- \\ m^- \\ d^- \end{bmatrix} \quad (25)$$

These relations have been verified experimentally using the nonlinearity of Figure 11. A comparison of the measured values of h, m, and d with the calculated values obtained by conventional least-squares analysis yields:

	h^+	m^+	d^+	h^-	m^-	d^-
Calculated	0.437	0	0.625	0.265	0	-0.156
Measured	0.456	-0.086	0.694	0.246	-0.056	-0.196

Use of these experimental data in the above matrix expression, yields estimates of the harmonic distortion with sinusoidal input. Measurements of the amplitude of these harmonic components were made with a wave analyzer, yielding the following comparison of harmonic amplitudes.

	a_o	a_1	a_3	b_2
Est. from Eq. (25)	0.211	0.525	-0.236	-0.060
Measured (absolute value)	0.230	0.545	0.243	0.053

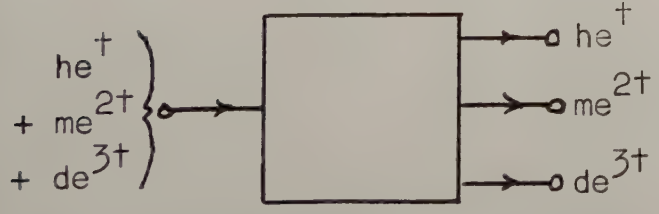


Fig. 1. An ideal exponential filter separates a signal into its exponential components.

Acknowledgment

This work was sponsored by the Air Force Cambridge Research Center, ARDC, under Contract AF 19(604)-1941, and by the Office of Naval Research under Contract Nonr-248(53). Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

References

1. L. A. Zadeh, "A General Theory of Linear Signal Transmission Systems", J. Franklin Inst. vol. 253, no. 4, April 1952.

2. W. H. Huggins, "Signal Theory", IRE Trans. CT3, no. 4, pp. 210-216, December 1956.

3. W. H. Huggins, "Representation and Analysis of Signals, Part I: The Use of Orthogonalized Exponentials", Report No. AFCRC TR-57-357, The Johns Hopkins University, Sept. 30, 1957.

4. D. C. Lai, "Representation and Analysis of Signals, Part II: An Orthonormal Filter for Exponential Waveforms", Report AFCRC TN-58-191, The Johns Hopkins University, June 15, 1958.

5. K. S. Miller, "Engineering Mathematics", pp. 178-179, Rinehart and Company, Inc., New York, 1957.

6. L. A. Zadeh, "On the Identification Problem", IRE Trans. CT-3, no. 4, pp. 277-281; December 1956.

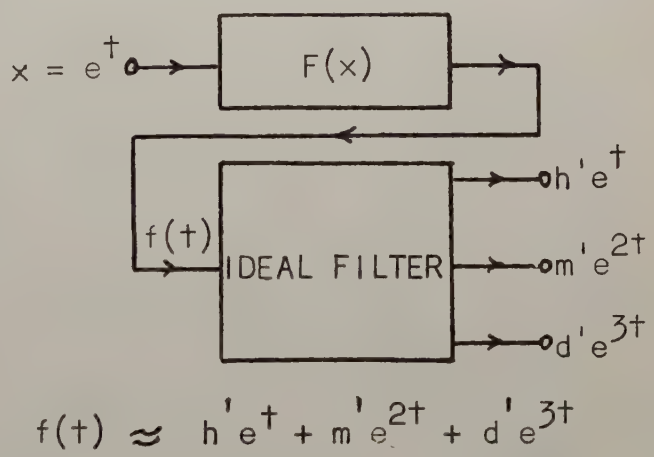


Fig. 2. Analysis of the harmonic exponential components produced by a nonlinear system by means of an ideal filter.

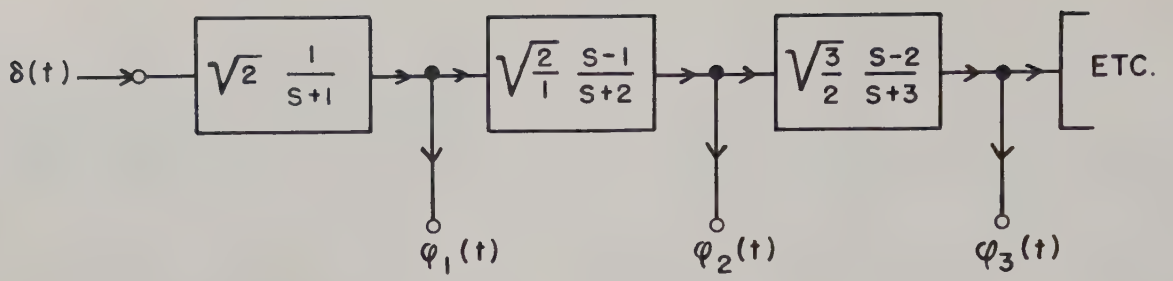


Fig. 3. An orthogonal filter.

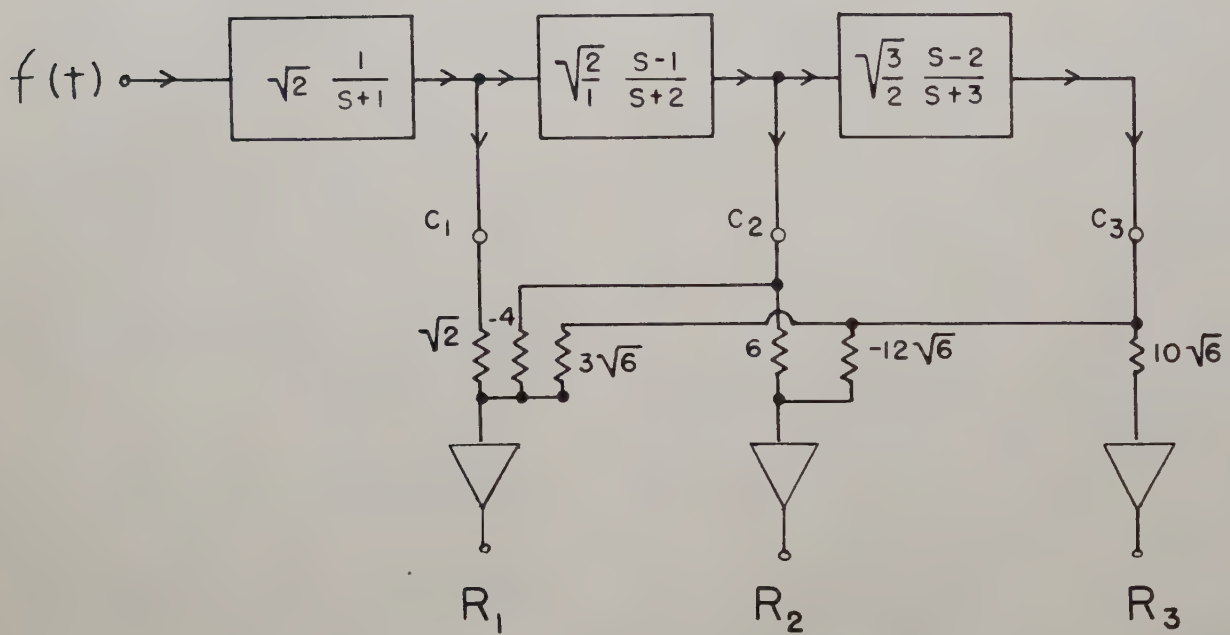


Fig. 4. Realization of an ideal filter by use of an orthogonal filter.

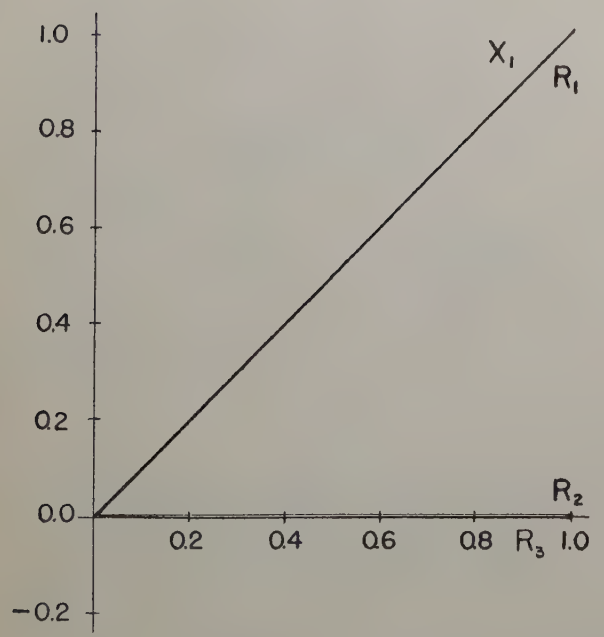


Fig. 5. The response of an ideal filter to an input $x_1 = e^t$.

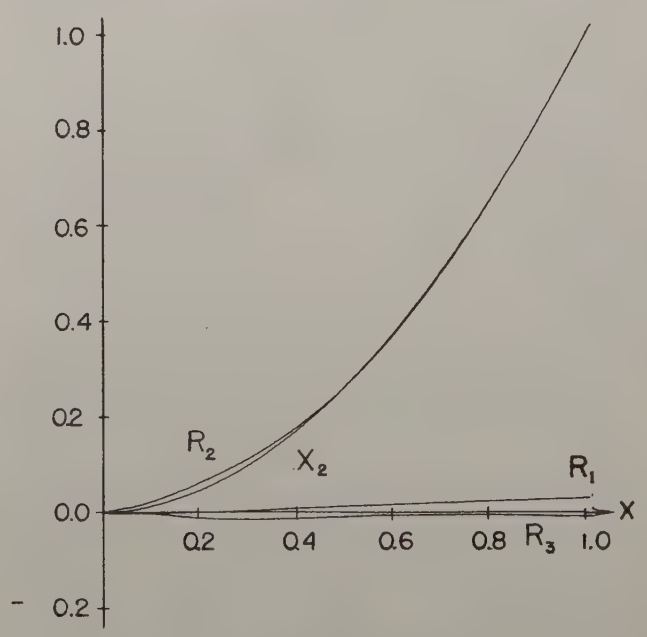


Fig. 6. The analysis of the output x_2 of a "squaring" device when driven by $x = e^t$.

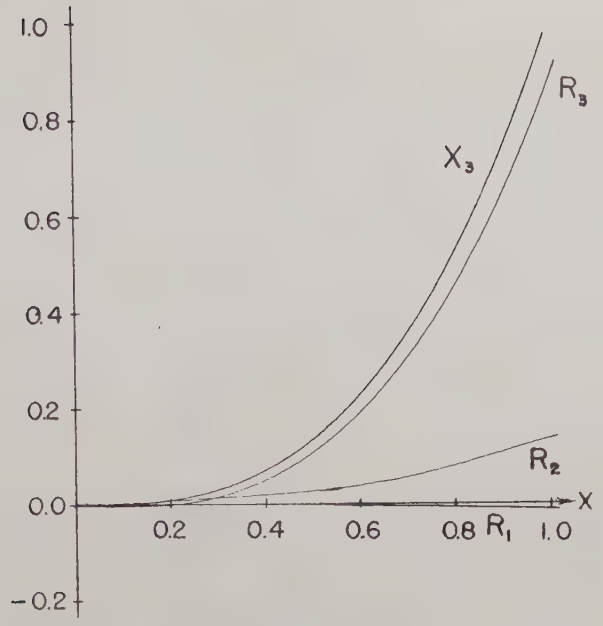


Fig. 7. The analysis of the output x_3 of a "cubing" device when driven by $x = e^t$.

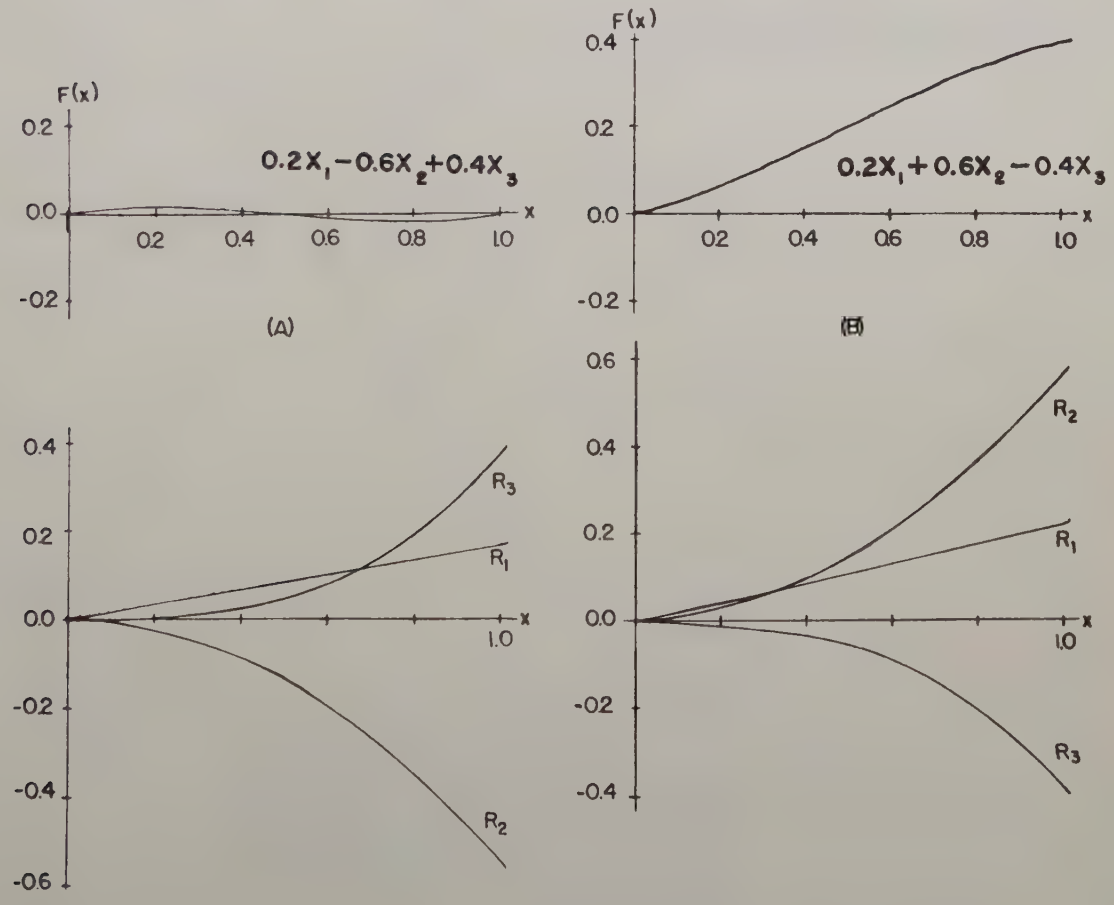


Fig. 8. The identification of two nonlinearities $F_A(x)$ and $F_B(x)$.

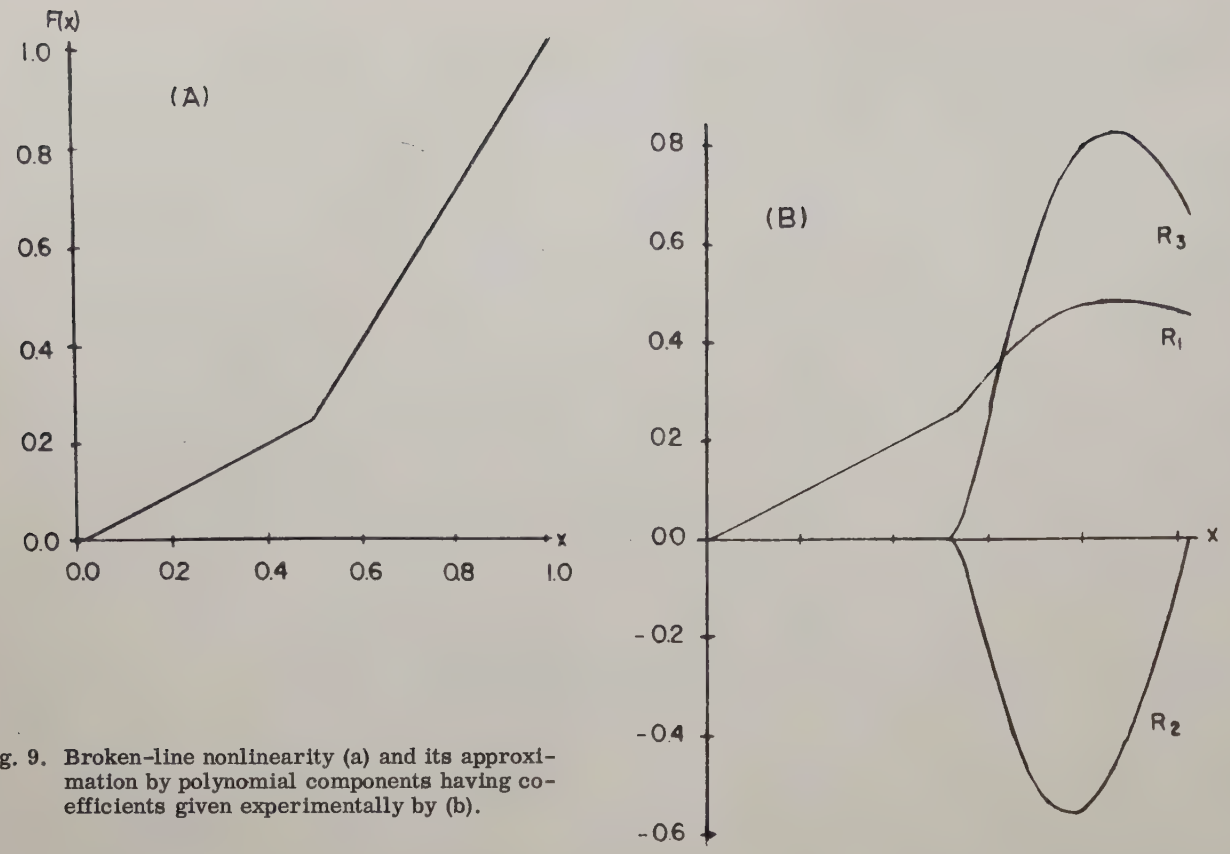


Fig. 9. Broken-line nonlinearity (a) and its approximation by polynomial components having coefficients given experimentally by (b).

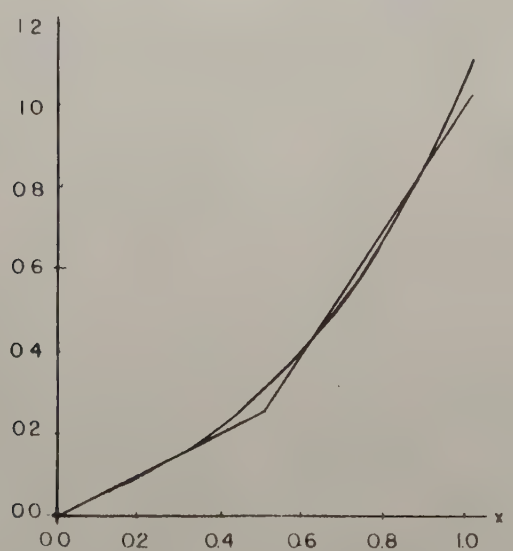


Fig. 10. The polynomial approximation to the broken line characteristic of Fig. 9, over the interval $0<x<1$ using $F(x) = 0.456x_1 - 0.086\ x_2 + 0.694\ x_3$ where the coefficients are the values of R_1, R_2 and R_3 obtained from Fig. 9(b) at $x = 1$.

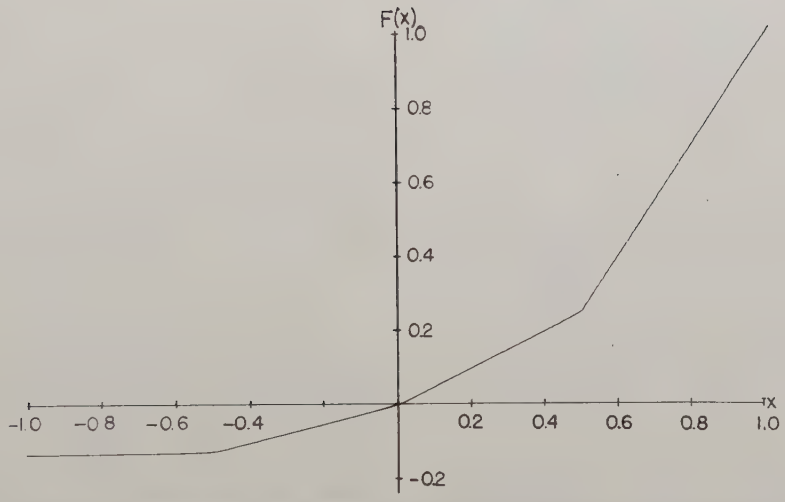


Fig. 11. Two-sided nonlinearity used for sinusoidal vs. exponential comparisons.

A PARAMETER TRACKING SERVO FOR ADAPTIVE CONTROL SYSTEMS

M. Margolis, Hughes Aircraft Company
and the University of California at Los Angeles

C. T. Leondes, University of California
at Los Angeles

Summary

This paper describes one very general approach to the design of adaptive control systems. The particular systems considered are process adaptive. The dynamic characteristics of the physical process are determined by the parameter tracking servo. The parameters thus determined are used to program the process' controller.

The parameter tracking servo is a closed loop self-adjusting system. It consists of the following elements; 1) the physical process, 2) the learning model, 3) the adjusting mechanism. The learning model and the physical process are subjected to the same input signals. Their outputs are compared and the resultant error is fed to the adjusting mechanism where some function of this error is used to adjust the parameters of the learning model.

The mechanism will continuously track the parameters of the physical process as they change with time in some unknown manner. The adjusting mechanism operates on an approximation to the method of steepest descent. These equations are derived for a first order process and the overall systems is analyzed. The equations describing the tracking servo's operation are both nonlinear and non-autonamous. System response as a function of input signal, gain, and error function are described analytically. Experimental results are included to demonstrate the validity of the analytic solutions.

Introduction

An adaptive control system is defined as a feedback control system intelligent enough to adjust its characteristics in a changing environment so as to operate in an optimum manner according to some specified criterion. The classical feedback amplifier which maintains almost constant gain in the face of variations in the gain of a particular stage may be classified as an adaptive system. In such a system, the adaptation that occurs is provided by changes in the feedback signal.

For more complex systems or systems whose characteristics change over a wide range simple passive adaptation techniques may not be sufficient.[1] We will restrict our attention to systems which are actively adaptive, and in particular to a type which has come to be known as process adaptive. A process adaptive system is one which determines the values of the significant parameters in the physical process and uses these values to program the controller according to the specified control laws.

The physical process must either be indeterminate or vary widely in its dynamic characteristics as its environment changes to justify the use of a process adaptive system. A particularly good example of such a process is a high performance interceptor aircraft whose dynamic characteristics change as a function of speed and altitude. The use of a process adaptive system then makes it possible to design the complete control system for a specified requirement without optimizing the controller parameters for the range of dynamic characteristic variation. The parameter tracking servo will determine these variations and present the "exact" values to the proper computing circuits for adjusting the parameters in the controller.

Emphasis is placed on the design of the parameter tracking servo. Methods for designing the controller programmer and the choice of the proper control laws are not discussed here.

The Learning Model Approach

The particular mechanization described in this paper is shown in Figure 1. A learning model whose characteristics are as nearly like that of the physical process as possible is used to program the pre-filter, feedback controller, and feedforward controller. The programming is done according to the appropriate control laws devised by the system designer.

The learning model is a computer mechanization, analog or digital, whose parameters are made available for use in the computing circuits

This research was supported in part by the United States Air Force under Contract No. AF 49(638)-438 monitored by the AF Office of Scientific Research of the Air Research and Development Command.

of the controller programmer. An example of a learning model with a simple analog mechanization is that represented by a linear differential equation.

$$\sum_i \alpha_i \frac{d^i y}{dt^i} = m(t) \qquad (1)$$

The parameters α_i are adjusted by the adjusting mechanism so that the behavior of the learning model is as much like the physical process as possible.

The Parameter Tracking Servo

The physical process is represented by a linear differential equation whose parameters a_i may be considered to vary with time in some unknown manner.

$$\sum_i a_i \frac{d^i z}{dt^i} = m(t) \qquad (2)$$

It is important to note that the parameters a_i are not considered random variables with an assumed mean, variance, and/or higher order moments.

The adjusting mechanism which is the heart of the parameter tracking servo then adjusts the parameter α_i to track the a_i as the a_i vary with time. The only information known by the adjusting mechanism is the error between the learning model and the physical process.

$$\epsilon = y - z \qquad (3)$$

The Method of Steepest Descent

The parameters α_i are adjusted according to the gradient of which the ith parameter adjusting equation is

$$\frac{d\alpha_i}{dt} = -k \frac{\partial f(\epsilon)}{\partial \alpha_i} \qquad (4)$$

If f(e) is chosen to be an even function whose minimum exists when the $\alpha_i = a_i$, then (4) will be satisfied. Any other values of α_i should find the gradients non-zero and of the proper sign so that the α_i are adjusted in the proper direction to approach the a_i.

The expression (4) is referred to as the equation of steepest descent. If the system is a multi-parameter stationary system, then the path of steepest descent in multi-dimensional Euclidean space is given by:

$$\frac{d\beta_i}{dt} = -c \frac{\partial f}{\partial \beta_i} \qquad (5)$$

where $\beta_1, \beta_2, \beta_3 \ldots \beta_i \ldots \beta_n$ are the n parameters. A rather thorough discussion of the method of steepest descent is given in reference 2. Each of the components of the gradient determine the rate at which its particular variable will change and the direction, positive or negative, that it takes.

To obtain a gradient, however, we must be able to describe a surface in the n-dimensional space. Unfortunately, f = f(e) is dependent on the type of forcing function to the learning model and physical process. Therefore, we are faced with a surface in n-space which is changing continuously as a function of time. Of far greater concern is the fact that a zero gradient may occur when the $\alpha_i \neq a_i$. A simple example exists when m(t) = 0. In this case y = z = ϵ = 0 and f(e) if it contains no constant term will be zero. If we ignore the analytical procedures and consider the system from an engineering point of view, then whenever m(t) = 0, there is no energy to the system and no possibility of any adjusting action. Therefore, there must be a forcing function for the tracking servo to operate.

Although no complete surface can be drawn in n-space to represent the function $f(\epsilon)$, we can think of a particular region at a particular instant and determine the change in $f(\epsilon)$ as a function of the several variables $\alpha_1, \alpha_2, \ldots \alpha_i \ldots \alpha_n$. Forcing the α_i to follow their respective components of a derived gradient will be shown to cause the proper tracking action.

Mechanization of the Gradient

The physical process that the learning model must track will be represented by a first order differential equation

$$\frac{dz}{dt} + az = m(t) \qquad (6)$$

The learning model is then described by a similar expression

$$\frac{dy}{dt} + \alpha y = m(t) \qquad (7)$$

Both systems are subject to the same forcing function, of course. Since only one parameter is being adjusted only one equation is required in the operation of the adjusting mechanism.

$$\frac{d\alpha}{dt} = -k \frac{\partial f(\epsilon)}{\partial \alpha} \qquad (8)$$

Obtaining the gradient of $f(\epsilon)$ with respect to the variable α for a system operating in real time is not possible. Some approximations are made after a suitable choice of $f(\epsilon)$ is found.

A simple even function for $f(\epsilon)$ is

$$f(\epsilon) = \epsilon^2 \tag{9}$$

another which may be tried is

$$f(\epsilon) = \int_{t_o}^{t} \epsilon^2 \, d\tau \tag{10}$$

It turns out that (10) leads to an unstable tracking servo with very small values of gain, k.

If (9) is substituted into (8), we have

$$\frac{d\alpha}{dt} = -2k \, \epsilon \, \frac{\partial \epsilon}{\partial \alpha} \tag{11}$$

But $\epsilon = y - z$ by (3) so that (11) becomes

$$\frac{d\alpha}{dt} = -2k \, (y - z) \, \frac{\partial y}{\partial \alpha} \tag{12}$$

Note that the $\partial z / \partial \alpha = 0$ since z is not a function of α.

The $\partial y / \partial \alpha$ can be approximated in the following ways; first, the approximation can be made using the first order terms of a Taylor series so that

$$\frac{\partial y}{\partial \alpha} \cong \frac{y(\alpha + \Delta\alpha, \, t) - y(\alpha, \, t)}{\Delta \alpha} \tag{13}$$

If an auxiliary equation of the following form is used

$$\frac{dy_1}{dt} + (\alpha + \Delta \alpha) \, y_1 = m(t) \tag{14}$$

then (13) becomes

$$\frac{\partial y}{\partial \alpha} \cong \frac{y_1 - y}{\Delta \alpha} \tag{15}$$

$\Delta\alpha$ is made a small but constant value in comparison to the range of values α is expected to cover. The system of equations describing the operation of the parameter tracking servo for a first order process are (6), (7), (12), and (14).

Another approximation of the gradient may be found by the following method.[3] For the present, consider the parameter α constant, then the ordinary differential equation (7) is exactly equivalent to the simple partial differential equation

$$\frac{\partial y}{\partial t} + \alpha y = m(t) \tag{16}$$

Taking the partial derivative with respect to α of (16)

$$\frac{\partial^2 y}{\partial \alpha \, \partial t} + \alpha \frac{\partial y}{\partial \alpha} + y = 0 \tag{17}$$

If α is still considered to be a constant valued parameter and the $\partial y / \partial \alpha = u$, (17) becomes

$$\frac{du}{dt} + \alpha u = -y \tag{18}$$

Thus far, Equations (16), (17) and (18) are exact. If α is permitted to vary, then (18) is only an approximate expression and u is only an approximation to the gradient $\partial y / \partial \alpha$. For slowly varying systems the approximation is very good. As the speed of response of the tracking servo is asked to increase, the value of u is less valuable. A second set of equations describing the performance of the parameter tracking servo will then be described by (6), (7), (12), and (18).

Performance of the Parameter Tracking Servo

The second mechanization of the gradient will be used in the analysis of the performance of the parameter tracking servo. The stability and dynamic response are examined for step inputs and sinusoidal inputs. Experimental verification in the form of analog simulation runs are also provided.

Step Inputs

The parameter tracking servo with a step input, M, is described by

$$\begin{aligned} \dot{z} + az &= m(t) = M \\ \dot{y} + \alpha y &= m(t) = M \\ \dot{u} + \alpha u &= -y \\ \dot{\alpha} &= -2k(y - z)u \end{aligned} \tag{19}$$

Equation (19) can be transformed to a standard form by the proper change of variable

$$\begin{aligned} z &= \frac{M}{a} + x_1 \\ y &= \frac{M}{a} + x_2 \\ u &= -\frac{M}{a^2} + x_3 \\ \alpha &= a + x_4 \end{aligned} \tag{20}$$

Equation (19) can then be represented in the form

$$\dot{x} = Ax + f(x) \tag{21}$$

where $\dot{x}$ and x are vectors whose components are the transformed members of the individual equations in (19), f(x) is the vector of non-linear terms and A is a square matrix of constant elements. A is given by

$$A = \begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & -a & 0 & -M/a \\ 0 & -1 & -a & M/a^2 \\ -2kM/a^2 & 2kM/a^2 & 0 & 0 \end{bmatrix} \qquad (22)$$

and f(x) by

$$f(x) = \begin{bmatrix} 0 \\ -x_2x_4 \\ -x_3x_4 \\ 2kx_1x_3 - 2kx_2x_3 \end{bmatrix} \qquad (23)$$

The determination of the stability of the null solution of the non-linear equation (21) is due to Liapunov[4] and Poincare.[5] Bellman states the following theorem and gives its proof in his book.[6]

Theorem 1. If a) Every solution of $v = A\,v$ approaches zero as $t \to \infty$
b) f(x) is continuous in some region about x = 0
c) $\|f(x)\| / \|x\| \to 0$ as $\|x\| \to 0$

Then x = 0 is a stable solution of (21)

Furthermore, every solution of (21) for which x(o) is sufficiently small approaches zero as $t \to \infty$.

$\|x\| = \sum_i |x_i|$ and $\|f(x)\| = \sum_i |f_i(x)|$ are spoken of as norms. If the characteristic roots of the A matrix all have negative real parts, then part (a) Theorem 1, is satisfied. The parameter tracking servo with m(t) = M has an A matrix whose characteristic roots are given by

$$(\lambda + a)^2 \left(\lambda^2 + a\lambda + \frac{2k\,M^2}{a^3}\right) = 0 \qquad (24)$$

$$\lambda = -a,\ -a,\ -a/2 \pm j\,a/2\sqrt{\frac{8k\,M^2}{a^5} - 1} \qquad (25)$$

As (25) indicates, the parameter tracking servo's linearized portion satisfies (a) of Theorem 1. On examining (23), f(x) is seen to be continuous about x = 0. And finally since f(x) is second order while x is first order, (c) of Theorem 1 is satisfied.

Through the use of Theorem 1, the tracking servo is seen to be asymptotically stable about the equilibrium point x = 0 which implies $\alpha = a$. There are several restrictions to the general application of the stability shown above which must be pointed out:

1) An autonamous system has been examined since
   a) only step inputs have been considered
   b) a is considered constant.
2) Stability in the small has been proved. The stability region about x = 0 has not been fully described.

Though these are serious restrictions which must be removed, the steady state stationary case must first be proved stable before any further discussion of stable tracking action is deemed worthwhile.

The dynamic response of the tracking action in the region about the point α near a can be found from the linear differential equation for x_4. x_4 is the variation of α from a. Using the linearized portion of (21) since f(x) is small near the null solution and solving for x_4

$$\ddot{x}_4 + a\dot{x}_4 + \frac{2k\,M^2}{a^3}\,x_4 = 0 \qquad (26)$$

and

$$x_4 = \frac{x(o)}{\omega}\, e^{-\frac{a}{2}t} \sin \omega t \qquad (27)$$

where

$$\omega = \frac{a}{2}\sqrt{\frac{8k\,M^2}{a^5} - 1}$$

Figure 2 gives the root location as a function of gain for the tracking system. It is seen to be a conventional root locus diagram for a second order system whose open loop poles lie on the real axis. The pole at the origin is due to the integrating action of the adjusting mechanism. The other pole is located at the root of the physical process -a.

Actually the variation equation pertains over a considerably larger region than was expected. Computer data for initial offsets of $\alpha = \alpha_o$ at $t = t_o$ illustrate behavior of an almost linear system. Figure 3 describes responses for an offset of $\alpha_o = 2.0$ and gain variations from 0.25 to 5.0. Figure 4 describes the result for varying initial offsets and a constant gain. In Figure 4 the non-linear response is evident. Note

however, that the terminal response is about the same.

It is important to note that the speed of response is directly dependent upon the value a. It should, of course, be expected that a faster physical process will give rise to a faster tracking servo. Figure 5 gives the root location for a 0.7 damped system with a unit step input and the value of k as a function of root location. As can be seen higher gains are required for faster systems. On examination of (26), the effective gain is modified by the value of a^{-3}. The faster the process, the larger the value of a and the smaller the effective gain $2kM^2/a^3$.

In all the equations describing the response of α, or more exactly its variation x_4, the gain k and the magnitude of the step M appear together as kM^2. This is really as it should be since the energy for operating the system is derived from the input, m(t). Without an input signal, m(t), as previously noted, the system would be dead. The signal level m(t) can be considered as the square root of the energy in the system and it may be desirable to vary the system gain k inversely as the $\sqrt{m(t)}$ to keep a desired response for the parameter tracking servo. In addition, since $\alpha \approx a$ is available in the learning model, the speed of response of the parameter tracking mechanism may be kept in a desired range by making k a function of a^3.

Sinusoidal Input Signals

One objective of the design of this parameter tracking servo is to obtain stable response no matter what the form of the input m(t). In fact a most happy situation would exist if α could track a with little or no error independent of the input signal waveform. Since no general theorems exist guaranteeing the asymptotic stability of x_4 for general bounded inputs m(t), it behooves us to attempt to study the effect of as general a class of inputs as possible. Sinusoidal inputs probably fit the category of a general input best. In fact, it is to be expected that a periodic waveform will give much more of a stability problem than an aperiodic waveform. There is always the problem of induced resonance effects.

It was previously observed that in the event of a step input, the system was always stable no matter how high the gain or how large the step, providing of course that the physical process is stable, $a > 0$. Is this necessarily true of an $m(t) = M_o \sin \omega t$? It turns out that the system is not only sensitive to the magnitude of the input, but also to the frequency of the sine wave.

The equations (19) with $m(t) = M_o \sin \omega t$ and

$$\begin{aligned} z &= x_1 \\ y &= x_2 \\ u &= x_3 \\ \alpha &= a + x_4 \end{aligned} \tag{28}$$

becomes

$$\begin{aligned} \dot{x}_1 &= -ax_1 + M_o \sin \omega t \\ \dot{x}_2 &= -ax_2 - x_2 x_3 + M_o \sin \omega t \\ \dot{x}_3 &= -x_2 - ax_3 - x_3 x_4 \\ \dot{x}_4 &= 2k\, x_1 x_3 - 2k\, x_2 x_3 \end{aligned} \tag{29}$$

These equations can be written in the following vector notation

$$\dot{x} = Ax + f(x) + M(t) \tag{30}$$

where the components of the vectors and matrix are as in (29).

If an attempt is made to solve (29) for x_4, the following differential equation obtains

$$\ddot{x}_4 + 2a\,\dot{x}_4 + x_4 x_4 - 2k\, x_2 x_3 x_4 = 2k\, x_2(x_2 - x_1) \tag{31}$$

In examining the stability about the equilibrium point $\alpha = a$ or $x_4 = 0$, second order terms may be neglected. (31) then becomes

$$\ddot{x}_4 + 2a\,\dot{x}_4 - 2k\, x_2 x_3 x_4 = 2k\, x_2(x_2 - x_1) \tag{32}$$

If x_4 is very close to zero, $x_2 = x_1$, and the RHS of (32) is very small. x_2 and x_3 can now be considered time varying quantities provided $x_4 \ll 1$. If m(t) is

$$m(t) = M_o \sin(\omega t + \psi) \tag{33}$$

then

$$x_2(t) = \frac{M_o}{(a^2 + \omega^2)^{1/2}} \sin \omega t \tag{34}$$

and

$$x_3(t) = - \frac{M_o}{a^2 + \omega^2} \sin(\omega t - \psi) \qquad (35)$$

where $\psi = \tan^{-1} \frac{\omega}{a}$.

$$x_2 x_3 = - \frac{M_o}{(a^2 + \omega^2)^{3/2}} \sin \omega t \sin(\omega t - \psi) \qquad (36)$$

$$x_2 x_3 = - \frac{M_o}{2(a^2 + \omega^2)^{3/2}} \left[\cos \psi - \cos(2\omega t - \psi)\right] \qquad (37)$$

where

$$\cos \psi = \frac{a}{(a^2 + \omega^2)^{1/2}}$$

on using (35) in (32) and letting RHS = 0, (32) becomes

$$\ddot{x}_4 + 2a \dot{x}_4 + \left\{ \frac{ak M_o^2}{(a^2 + \omega^2)^2} - \frac{k M_o^2}{(a^2 + \omega^2)^{3/2}} \cos(2\omega t - \psi) \right\} x_4 = 0 \qquad (38)$$

(38) will be recognized as a Mathieu equation with a damping term.[7,8] It can be transformed to a standard Mathieu form by the following transformations

$$\tau = \omega t \qquad (39)$$

and

$$x = e^{-(\frac{a}{\omega})\tau} y \qquad (40)$$

(38) then becomes

$$\frac{d^2 y}{d\tau^2} + \left\{ a_1 - 2q_1 \cos(2\tau - \psi) \right\} y = 0 \qquad (41)$$

where

$$a_1 = \frac{ak M_o^2}{\omega^2(a^2 + \omega^2)^2} - \frac{a^2}{\omega^2} \qquad (42)$$

$$q_1 = \frac{k M_o^2}{2\omega^2(a^2 + \omega^2)^{3/2}} \qquad (43)$$

The solution of (41), the standard Mathieu equation, can be stable, periodic, or unstable. We are particularly concerned if (41) results in an unstable solution. The form of the unstable solution of (41) is

$$y = e^{\mu\tau} \phi(\tau) + e^{-\mu\tau} \phi(-\tau) \qquad (44)$$

where μ is a positive constant determined by a_1 and q_1 and $\phi(\tau)$ is a periodic function. The solution of (38) is actually our true concern and it is given by

$$x = e^{-\frac{a}{\omega}\tau} \, y = e^{-(\frac{a}{\omega} - \mu)\tau} \phi(\tau) + e^{-(\frac{a}{\omega} + \mu)\tau} \phi(-\tau) \qquad (45)$$

As long as $\frac{a}{\omega} > \mu$ the system has a stable singular point $\alpha = a$. μ is related to a_1 and q_1 in a very difficult way. Figure 6 describes the regions of stability and instability of (41). Figures 7, 8 and 9 show a detailed set of iso - μ lines for the first three unstable regions.[9]

A particular case was taken to illustrate the validity of approximations made above in examining the equilibrium of the system subject to sinusoidal signals. The following conditions were chosen

$$a = 1.0$$

$$m(t) = \sqrt{5} \sin 2t$$

for which $\frac{a}{\omega} > |\mu|$, $1/2 > |\mu|$ is the stability condition. Figure 7 shows that for the following gains the value of μ will be as shown in Table I.

TABLE I

k	a_1	q_1	μ
5	0.0	.28	stable
10	0.25	.56	stable
15	0.50	.84	-0.4 stable
17.5	0.625	.98	-0.49 borderline
20	0.75	1.12	-0.56 unstable

Experimental verification of the predicted results was obtained with the use of an analogue computer. Figure 10 describes the response to gains of k = 10 and 20. Periodic oscillations may be observed in the computer runs even in the stable case, k = 10. It should be recalled, however, that the RHS of (38) is not necessarily exactly zero if $x_2 \neq x_1$ and that there are non-linear terms which have been ignored on the basis that the response x_4 remained in a small region about $x_4 = 0$.

A completely satisfactory response can be expected if a suitable choice of k is made. Figure 11 describes the response of α with an initial offset and a sinusoidal input signal for gains of k = 1 and 4. In these runs $\omega = 2.0$. A set of runs was also taken for various frequencies of the sinusoidal input signal. Figure 12 describes the response of α for an initial offset. In all these cases the amplitude of the sinusoidal input signal was so chosen that the magnitude of the sinusoidal output of the process was unity. The same effective gain $kM_o^2/(a^2+\omega^2)$ then pertains to all the runs in Figure 12.

Tracking Capability of the Learning Model

The ability of the parameter α to follow the process parameter a was tested under the following conditions. Both the model and process were subjected to a constant step input, $m(t) = M = 1$. The parameter in the process was made to vary in a sinusoidal manner with time about a mean value so that

$$a(t) = a_o + a_1 \sin \omega t \qquad (46)$$

where $a_o = 2.0$ and $a_1 = 0.25$. The analogue computer was mechanized to perform these tests. Although the parameter tracking servo is non-linear in its operation, the system is operating in a small enough region about a_o to be considered quasi-linear. Under these conditions, therefore, a frequency response test was made.

For the gains k = 8, 20, and 40, Figure 13 gives the frequency response in db. The variation equation (26) for $a = a_o$ and the gain as above give a damping ratio $\xi = .7$, .45 and .316, respectively. Curves for a linear stationary parameter system are drawn in Figure 13 for comparison.

The ability to track a(t) in what is really a positional tracking servo and not a velocity tracking system is dependent on the root location of the variational equation (26) for $a = a_o$. The higher the gain the greater the bandwidth and the better the tracking capability.

Conclusions

A particular form for a process adaptive control system has been suggested. This form makes use of what we call a Parameter Tracking Servo. The parameters to be tracked are the coefficients of the physical process differential equation. The tracking servo utilizes a learning model whose parameters are adjusted in such a manner that the model and process are as much alike in their dynamic behavior as possible. The criterion for similar behavior of the model and process was taken to be the minimum value of ϵ^2.

The ability of a first order model to track a first order process has been demonstrated. If the over-all feedback control system were to have a bang-bang actuator, step input signals would provide a sufficient test of the tracking servo's capability. In the more general case, for quasi-linear operation, the input signals will either be damped sinusoids or damped exponentials. For this case, the input signals were made sinusoids and under these rather severe conditions the tracking servos capabilities have been analyzed.

This work is but a start on what is felt to be a very general approach to the problem of process adaptive control systems. A report of research extending this work to higher order systems and to complete feedback control systems will be presented in the near future.[10]

Acknowledgements

The research reported in this paper was performed while the first named author was the recipient of a Hughes Staff Doctoral Fellowship. This research was supported in part by the United States Air Force under Contract No. AF 49(638)-438 monitored by the AF Office of Scientific Research of the Air Research and Development Command.

Dr. Ed Deland of the Rand Corporation was most helpful in making available the analogue computing facility of his organization for the experiments reported herein.

Figure 6 with some modification was taken from reference 7, page 114; and the basis for Figures 7, 8 and 9 can be found in reference 9, pages 9 and 10.

References

1. J. A. Aseltine, A. R. Mancini, and C. W. Sarture, "A Survey of Adaptive Control Systems," IRE Transactions on Automatic Controls, PGAC-6, December 1958, pp 102-109.

2. Beckenbach, Modern Mathematics for the Engineer, chapter 18 by C. G. Tompkins, McGraw-Hill Book Co., New York 1956.

3. Unpublished notes on "An Abridged Simulation Technique for Problems Involving Random Noise Signals," by H. F. Meissinger, Hughes Aircraft Company, Systems Development Laboratories, December 1957.

4. A. Liapunov, "Problem general de la stabilite du mouvement," Ann. Fac. Sci. Univ. Toulouse, Vol. 9 (1907), pp 203-475. Reprinted in the Annals of Mathematics Studies, 1947.

5. H. Poincare, Methods nouvelles de la mecanique celeste, Vols. I, III, Gauthier-Villars and Cie, Paris, 1892.

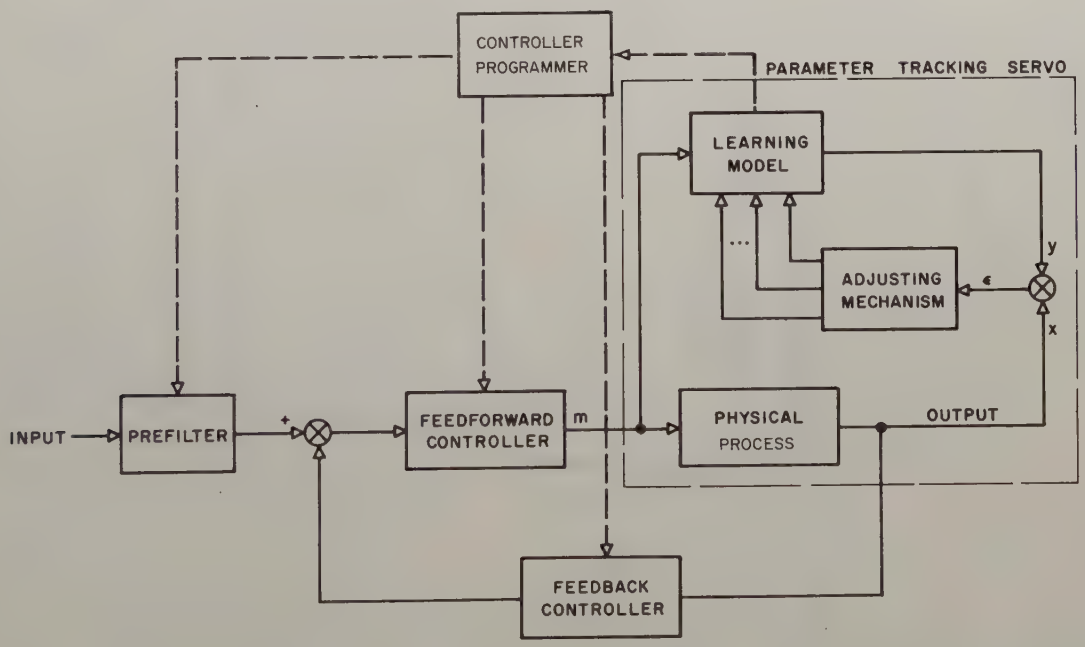
6. R. Bellman, Stability Theory of Differential Equations, McGraw-Hill Book Co., New York, 1953, pp 80-82.

7. N. W. McLachlan, Ordinary Nonlinear Differential Equations in Engineering and Physical Sciences, Oxford University Press, London, 1950, pp 113-119.

8. N. W. McLachlan, Theory and Application of Mathieu Functions, Oxford University Press, London, 1947.

9. C. Hayashi, Forced Oscillation in Nonlinear Systems, Nippon Printing and Publishing Company, Ltd., Osaka, Japan, 1953, Chapter I.

10. M. Margolis and C. T. Leondes, "On the Theory of Adaptive Control Systems, the Learning Model Approach," to be given at the forthcoming Congress of the International Federation for Automatic Control in Moscow, USSR, June 1960.



AN ADAPTIVE CONTROL SYSTEM USING A PARAMETER TRACKING SERVO TO PROGRAM THE PREFILTER, FEEDFORWARD AND FEEDBACK CONTROLLERS

FIGURE 1

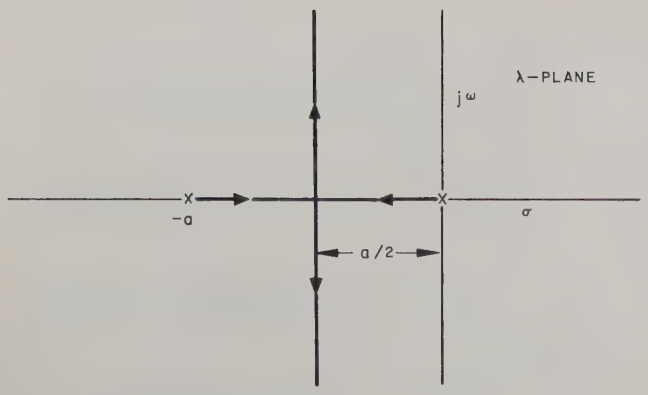


FIGURE 2 ROOT LOCATION AS GAIN k IS INCREASED

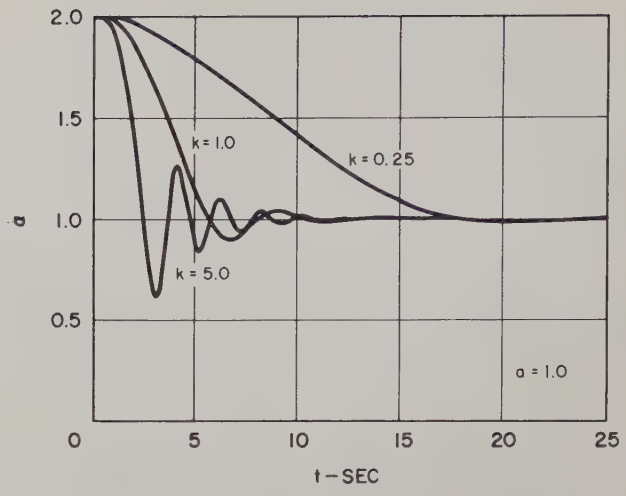
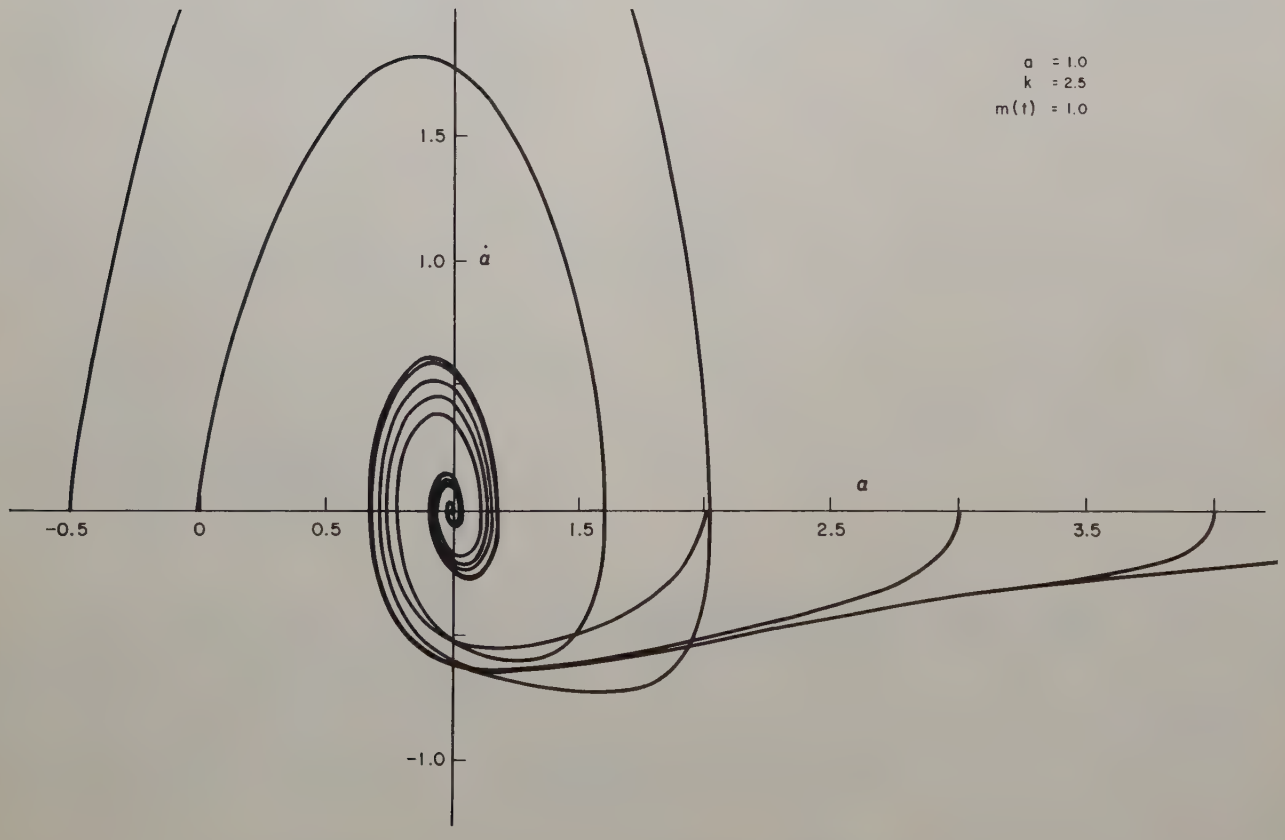


FIG. 3 THE RESPONSE OF α WITH AN INITIAL OFFSET AND m (t) = 1.0.



α

FIGURE 4 THE RESPONSE OF α WITH VARIOUS INITIAL OFFSETS

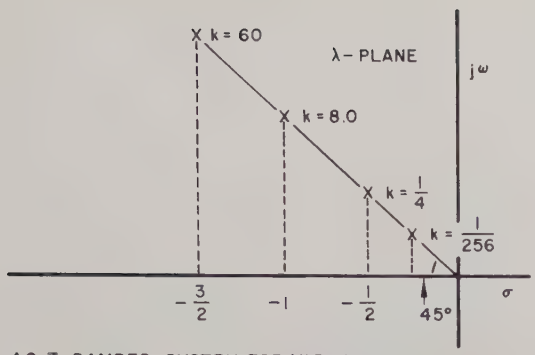


FIGURE 5 A0.7 DAMPED SYSTEM FOR VARYING a WITH A UNIT STEP INPUT

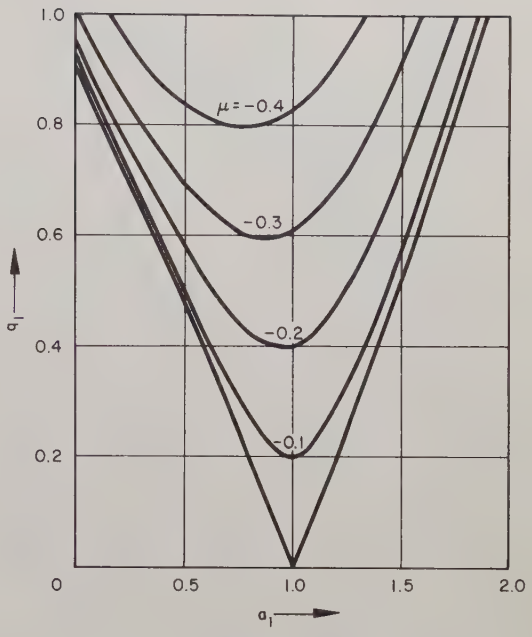


FIG. 7 ISO-μ CURVES IN THE FIRST UNSTABLE REGION.

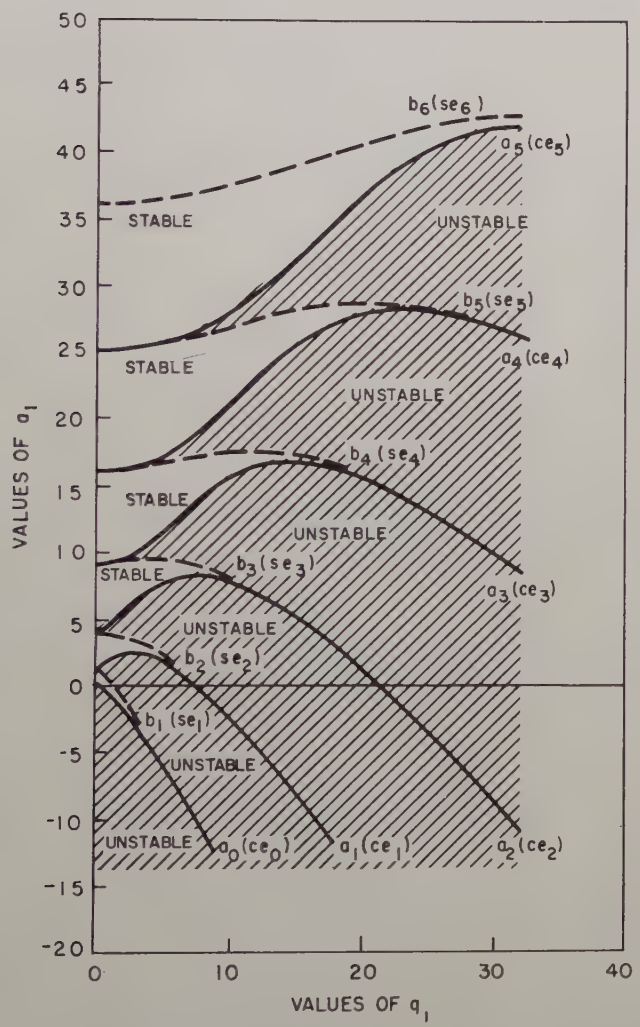


FIGURE 6 STABILITY CHART FOR THE MATHIEU EQUATION
$y'' + (a_1 - 2q_1 \cos 2\tau)\, y = 0$

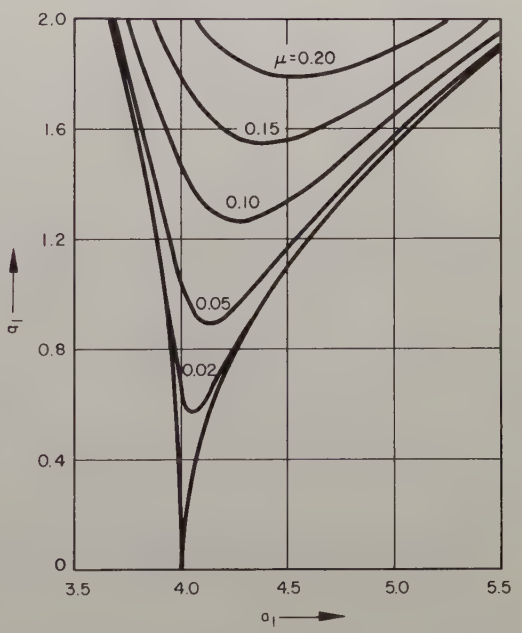


FIG. 8 ISO-μ CURVES IN THE SECOND UNSTABLE REGION.

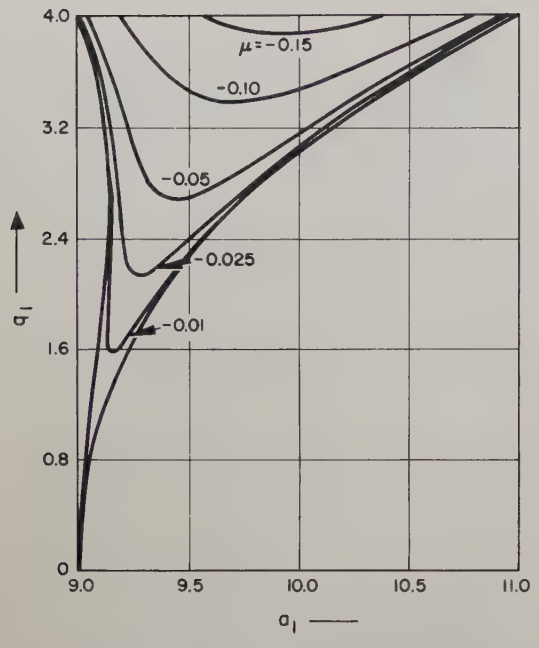


FIG. 9. ISO-μ CURVES IN THE THIRD UNSTABLE REGION

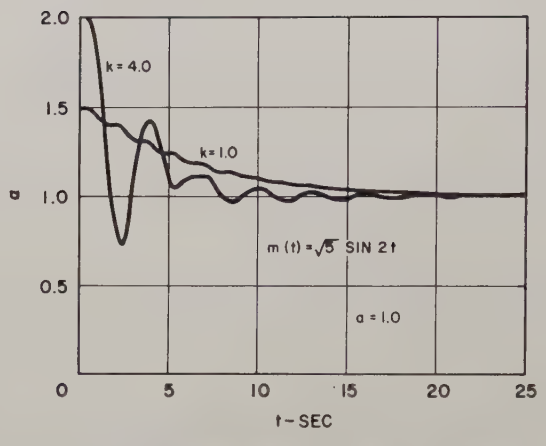


FIG. 11 THE RESPONSE OF a WITH AN INITIAL OFFSET AND A SINUSOIDAL INPUT SIGNAL.

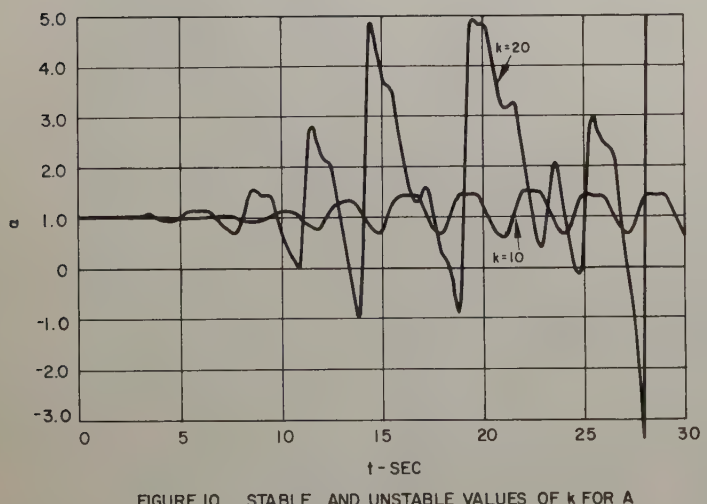


FIGURE 10 STABLE AND UNSTABLE VALUES OF k FOR A SINUSOIDAL INPUT SIGNAL

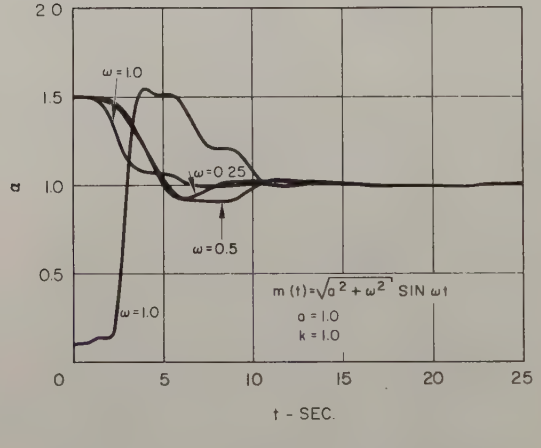


FIG. 12. THE RESPONSE OF a WITH AN INITIAL OFFSET AND A SINUSOIDAL INPUT SIGNAL FOR VARIOUS FREQUENCIES OF THE SINUSOID.

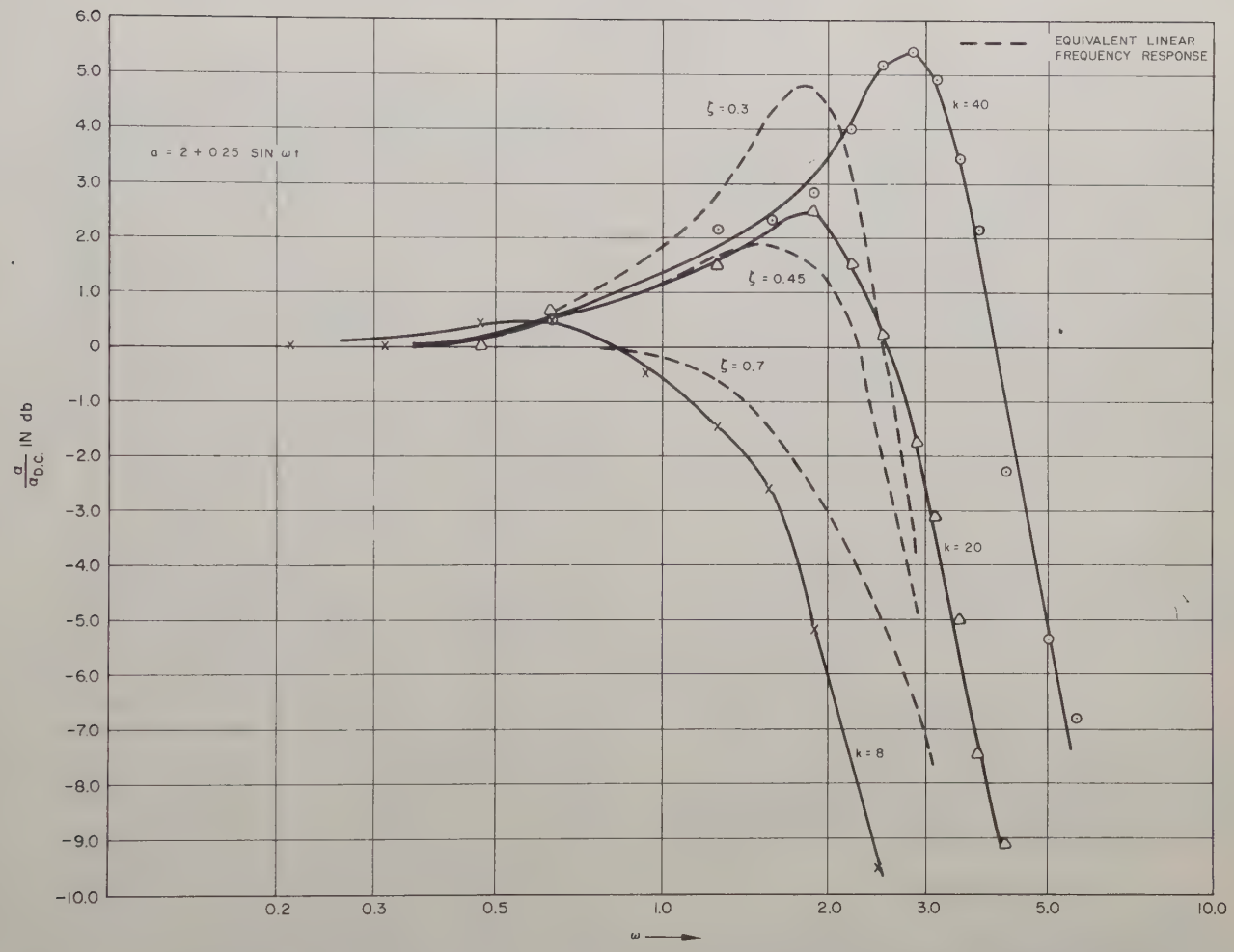


FIG. 13. THE RESPONSE OF a TO A SINUSOIDAL VARIATION IN A PLOTTED AS A FREQUENCY RESPONSE.

MAXIMUM EFFORT CONTROL FOR OSCILLATORY ELEMENT

Harold K. Knudsen
Department of Electrical Engineering
University of California
Berkeley, California

Summary

Maximum effort control is a method of achieving deadbeat response for a step input to an undamped second order element (two poles on the jω axis of the s-plane) which is preceded by a saturating amplifier. This method of control supplies the maximum available energy to the element being controlled, by driving the amplifier to saturation whenever an error is present. The realization of a maximum effort control system for an oscillatory element is found through a phase plane analysis of the equations of motion of the oscillatory element. The control system topology is also found by the analysis of a phasor representation of the transients introduced in the oscillatory element by the output of the saturating amplifier.

The system was constructed to compare the experimental responses of step inputs and load disturbances, to the responses obtained from an idealized mathematical model of the system. The adaptability of the control system was tested by using it to control a damped oscillatory element.

The system described will be an aid in the design of the control system which will give optimum response for random inputs to a undamped second order element which is preceded by a saturating amplifier. It also provides a method of control for systems in which it is impossible to damp poles near the jω axis.

Introduction

Maximum effort control of a system which has saturable components is more effective than linear control of the same system. When linear control is used, any saturation that occurs during system operation reduces loop gain. Reduced loop gain degrades system performance by decreasing the accuracy and speed of response. In contrast, maximum effort control drives a component to saturation whenever any error exists. This mode of operation utilizes the maximum available power to correct errors in the system output, thus giving fast response, while the accuracy of the output is determined by the open loop gain which exists when there is no saturation. The design of a maximum effort control system is essentially the design of a computer which will drive the saturating device in such a manner that error is reduced to zero in the minimum possible time.

The design of a computer for the control of an undamped element is approached through a study of the step response of the element. This is made possible by assuming the amplifier (the saturating device) has infinite gain. With infinite gain in the linear region of the amplifier, the output of the computer (the control function)[1] need only be a function that has the sign of the amplifier output which should be applied to the oscillatory element. (The amplifier output will have only two levels, positive saturation, and negative saturation.) The problem in the design of the computer now becomes one of finding the sign and duration of the steps of force which are applied to the oscillatory element. Three steps of force are necessary to obtain deadbeat response. For example, if a small positive step input is applied to the system, the computer output should first be positive, then after a short duration of time negative, and after another duration, positive again. The resultant amplifier output first forces the output of the oscillatory element to follow the system input, then subtracts energy from the element to prevent overshoot, and finally provides the necessary output level to make the error zero. Since the oscillatory part of the step response of the undamped element is sinusoidal, it can be represented by a phasor. If the three phasors representing the effect of the three steps of force applied by the amplifier have the proper timing their sum will equal zero. This represents deadbeat response. The system switch curve is found by relating the timing of the steps of force to error and error derivative. The locus of points in the phase plane at which the force applied to the oscillatory element is reversed is called the switch curve. The switch curve is also found through a study of optimum system operation in the phase plane.

The design of the computer is found through an interpretation of the shape of a trajectory in the phase plane. A trajectory is a representation of the equation of motion of the oscillatory element in terms of error and error derivative.

The system was set up on an analogue computer to observe the effects of the idealizations used to simplify the system design. It was found that low gain in the saturable amplifier degrades system performance. To improve the system performance, the effective gain of the amplifier was increased by the use of positive feedback. The system response was also observed when the oscillatory element was critically damped.

System Description

Configuration

The control system is built around two

unalterable components, a saturating amplifier, and an oscillatory element. The oscillatory element has a pair of poles on the jω axis of the s-plane and unity d-c gain. The control system is a computer which drives the saturating amplifier. The inputs to the computer are the system input (r) and the system output (c). A block diagram of the system is shown in Figure 1.

Response

The system is designed to give deadbeat response for a step input, and to reduce the error caused by it to zero in the minimum possible time. Deadbeat response occurs when the error, and all its derivatives become zero at the same time.

Idealizations

In the design of the computer, the gain of the saturating amplifier is considered infinite in its linear zone. This limits the output of the amplifier to its saturation limits. Also, it is assumed that the saturation levels are identical. The saturation levels are normalized to +1 and -1.

Computer Design

Computer Output

A method of determining the computer output is to find an input to the oscillatory element that will give a deadbeat response. This response must have the same final value as the system input to make the final value of the error zero. This is necessary because the d-c gain of the oscillatory element is unity. Figure 2 illustrates the output the saturating amplifier must have for a given step input. The amplifier output required when there is no error ($t<0$, $t>t_1 + t_2$) is obtained by a very high frequency oscillation of the computer output, which drives the amplifier between negative and positive saturation. The d-c level of the amplifier output is the average value of this oscillation. The step durations, t_1 and t_2, are determined from a phasor representation of the transients produced in the oscillatory element by the amplifier output.

Phasor Solution of Switching Times

Referring to Figure 2, it is seen that the amplifier output consists of three steps. Since the oscillatory element is linear, the output of the element is the sum of the three step responses. The magnitude of the first step is (1 - a), the magnitude of the second is 2, and the magnitude of the third is (1 + b).

Response to step 1:

$$1 - a - (1 - a) \cos \omega_n t \tag{1}$$

Response to step 2:

$$-2 + 2 \cos \omega_n (t - t_1) \tag{2}$$

Response to step 3:

$$1 + b - (1 + b) \cos \omega_n (t - t_1 - t_2) \tag{3}$$

The sum of the constant terms equal the step input (b - a). The sum of the sinusoidal parts of the response must equal zero for a deadbeat response. This condition can be satisfied by choosing the proper values for t_1 and t_2. Figure 3 (a) shows a phasor representation of the sinusoidal parts of the step responses. Figure 3 (b) shows the vector addition of these phasors, with t_1 and t_2 (θ_1 and θ_2) chosen to produce a sum of zero. Since the angle θ_1 is directly proportional to the first switching time t_1, this time can be found from the phasor diagram and interpreted in terms of error and error derivative. The locus of points in the phase plane at which the first reversal of the amplifier output takes place is called a switch curve.

Switch Curve from Phasors

From Figure 3 (b):

$$1 - a + (1 + b) \cos \theta_2 = 2 \cos \theta_1 \tag{4}$$

$$2 \sin \theta_1 = (1 + b) \sin \theta_2 \tag{5}$$

Solving for θ_1 by squaring equations (4) and (5) and combining them to eliminate θ_2:

$$\cos \theta_1 = \frac{(1 - a)^2 + 4 - (1 + b)^2}{4 (1 - a)} \tag{6}$$

To convert the first switching time into values of error and error derivative, the error and error derivative at t_1 must be computed. At time t_1:

$$e = r - c \tag{7}$$

where: r = system input, and c = system output

$$r = b \tag{8}$$

$$c = 1 - (1 - a) \cos \theta_1 \tag{9}$$

$$e = b - 1 + (1 - a) \cos \theta_1 = b - 1 + (1 - a) \cos \omega_n t_1 \tag{10}$$

$$\frac{\dot{e}}{\omega_n} = (a - 1) \sin \theta_1 \tag{11}$$

Substitute equation (6) into equation (10) and simplify to obtain

$$e = \frac{2b - 2a + a^2 - b^2}{4} \tag{12}$$

Substitute equation (6) into equation (11) to obtain

$$\frac{\dot{e}}{\omega_n} = (a - 1) \sin \cos^{-1} \frac{(1 - a)^2 + 4 - (1 + b)^2}{4 (1 - a)} \tag{13a}$$

or

$$\frac{\dot{e}}{\omega_n} = - \frac{\sqrt{16 (1 - a)^2 - \left[(1 - a)^2 + 4 - (1+6)^2\right]^2}}{4} \tag{13b}$$

Eliminate the variable a between equation (12) and equation (13b) and solve for $\frac{\dot{e}}{\omega_n}$ to obtain

$$\frac{\dot{e}}{\omega_n} = - \sqrt{2e + 2be - e^2} \tag{14}$$

Equation (14) is the solution for the switch curve for positive error at t = 0+. A similar derivation can be used for the case with negative error at t = 0+. For negative initial error:

$$\frac{\dot{e}}{\omega_n} = \sqrt{- 2e - 2be - e^2} \tag{15}$$

A general form of equations (14) and (15) which is valid for either positive or negative initial error can be found by inspection. Replacing b by r, the general form is:

$$\frac{\dot{e}}{\omega_n} = - \frac{e}{|e|} \sqrt{2 |e| + 2 er - e^2} \tag{16}$$

Equation (16) is a solution for the switch curve which is valid when the system input does not exceed the saturation levels of the amplifier, and when the quantity under the radical remains positive. The switch curve is recognized to be two semicircles of varying radii in a phase plane with error and normalized error derivative

$$\left(\frac{\dot{e}}{\omega_n}\right)$$

for axes. The effect of the system input on the radii of the semicircles is shown in Figure 4. The phasor approach gives a switch curve which is correct for step inputs; however, load disturbances may create errors of such magnitude that the system is forced to operate outside of the regions with a semicircular switch curve. Extensions of the switch curve to this region are found through a study of the system response in the phase plane.

System Response in the Phase Plane

Description of Phase Plane. The coordinates of the phase plane, used for the computer design, are error and normalized error derivative. The trajectories followed by the equations of motion of an undamped oscillatory element are circles centered at $\dot{e}$ equal zero, and e equal to the d-c level of the error. (The d-c level of the error is equal to the system input minus the amplifier output.) The direction of movement of a point on the trajectory is clockwise, with the angle formed by the moving point, the center of the circle, and some reference point, proportional to time. The radius of the trajectory is dependent on the initial conditions (e, $\dot{e}$ at t = 0).

Optimum System Operation. Bushaw[2] has proved that for certain initial conditions, a maximum effort control system containing an oscillatory second order element can have its error and error derivative reduced to zero in two trajectories (minimum time possible). Other initial conditions may require three or more conditions. All initial conditions which produce trajectories which cross the semicircular regions of the switch curve can be reduced to zero in two trajectories (one reversal of the amplifier output).

Phase Plane Approach to the Switch Curve. The mathematical solution for the switch curve can be justified by applying the criterion of optimum operation (two trajectories) to the system response. The method of justification chosen is to derive a switch curve that gives optimum response by observing the response of the oscillatory element to step inputs, and comparing this switch curve with that obtained mathematically.

The origin of the phase plane is a point at which error and error derivative are zero. This condition is fulfilled when the input equals the output (r = c), and the output derivative equals zero ($\dot{c}$ = 0). (For r(t) a step input.) Therefore the output of the amplifier must equal the system input for the error to be zero. (The oscillatory element has unity d-c gain.)

During the system response to a particular step input, only two levels of d-c error are available. They are the input signal minus the level of positive saturation (r - 1), and the input signal minus the level of negative saturation (r + 1). Therefore, all trajectories will be centered at either e = 1 + r, or e = -1 + r. (Trajectories are centered on the error axis at the d-c level of the error.) The only mode of operation which results in zero error and zero error derivative at the end of two trajectories, is one in which the second trajectory passes through the origin of the phase plane. Since the centers of the trajectories are known, and a point on the second trajectory is known, the second trajectories are defined. The only way to enter the second trajectory is to reverse the amplifier output as the first trajectory crosses it; thus, the second trajectory is also the switch curve. The switch curve is still not completely defined. Still to be found are, the location of the switch curve when the first trajectory does not intersect the trajectory which passes through the origin, and the portion of the trajectory which passes through the origin that should be used as a switch curve. The switch curve is chosen so

that minimum time is used in reaching the origin. The construction in Figure 5 shows that the parts of the second trajectories in the second and fourth quadrants should be used as switch curves. A similar construction can be used to show that the switch curve follows the error axis when the first trajectory does not intersect with a trajectory which passes through the origin.

This switch curve is identical to the one obtained through the phasor presentation of system action. In addition, through the phase plane study, the switch curve is defined for the entire phase plane. The design of the computer is found from an examination of the switch curve and its equation.

Realization of Computer

Consider the phase plane to be split into two sections by the switch curve. If the amplifier output is positive in the upper portion of the plane, and negative in the lower portion of the plane, the trajectories from any initial value will converge to the origin of the plane. To create these outputs, the computer must provide a positive signal when the input, error, and error derivative signals indicate the trajectory is in the upper portion of the plane, and a negative signal when the trajectory is in the lower portion of the plane. As an example of system operation, assume a step input identical to that of Figure 2 is introduced to the computer. This initial condition places the first trajectory in the upper portion of the phase plane; therefore, the amplifier output is positive. After a duration of t_1, the trajectory intersects the switch curve, the amplifier output changes sign, and the second trajectory is entered. The second trajectory reaches the origin at $t = t_1 + t_2$. At the origin, the amplifier oscillates between its saturation limits at a very high frequency. The average value of this oscillation is equal to the system input.

To obtain the computer design for the semicircular regions of the switch curve, the equation for the switch curve (16) is modified as follows:

Multiply equation (16) by its absolute magnitude to obtain

$$\frac{\dot{e}|\dot{e}|}{\omega_n^2} = -\frac{e}{|e|}(2|e| + 2er - e^2) \qquad (17)$$

expand equation (17) and rearrange to obtain

$$\frac{\dot{e}|\dot{e}|}{\omega_n^2} + 2e + \frac{e}{|e|}(2er - e^2) = 0 \qquad (18)$$

A device performing the operations indicated by the left side of equation (18) is used as a computer because the left half of equation (18) is positive in the portion below it. This computer produces a switch curve which is composed of two semicircles, and the dotted extensions to them shown in Figure 4. The proper switch curve outside of the semicircular regions is the error axis. To force the computer to follow this portion of the switch curve, the term

$$\frac{e}{|e|}(2er - e^2) + 2e$$

is made equal to zero by limiting the output (c). A block diagram of the computer is shown in Figure 5. In the computer, the term

$$\frac{e}{|e|}(2er - e^2)$$

is replaced by $|r - c|(r + c)$ as:

$$\frac{e}{|e|}(2er - e^2) = |r - c|(r + c) \qquad (19)$$

The system input (r) is clipped to the saturation levels of the amplifier because it is impossible to make the d-c level of the oscillatory element exceed these limits. The variable level clipper in the feedback path from the system output (c) forces the switch curve to follow the error axis outside of the regions where the switch curve is semicircular.

System Operation

The system was set up on an analogue computer. This made it possible to study the effects of non-infinite gain of the linear zone of the saturating amplifier. The responses to step inputs and load disturbances were observed.

Effect of Low Gain. When the gain of the saturating amplifier is not infinite, forces (amplifier outputs) smaller than the saturation levels of the amplifier occur when trajectories near the switch curve. This occurs because the computer output goes through zero as the trajectory crosses the switch curve. As a trajectory approaches the switch curve, a steadily decreasing force is applied to the oscillatory element once the amplifier is no longer saturated. When the trajectory crosses the switch curve, the force is reversed and increases as the trajectory moves away from the switch curve. The trajectory becomes concentric to the switch curve when the amplifier saturates. Since the trajectory does not coincide with the switch curve, it does not pass through the origin. The result is degraded system performance. Figure 7 illustrates the effect of gain on the system performance in the phase plane. Figure 8 shows the system response to a step input with an amplifier gain of five. These responses show that high gain gives superior performance. To improve system performance when the amplifier gain was low, positive feedback was placed around the saturating amplifier is shown in the phase plane in Figure 7, and versus time in Figure 9(b).

Response to Load Disturbance. The system response was also observed for step load disturbances which were larger than the available error

correcting forces. In these cases, the system requires more than two trajectories to remove the energy due to the load disturbance from the oscillatory element. The load disturbance response is shown in the phase plane in Figure 10(a) and versus time in Figure 10(b).

Application of the Computer to a Damped Oscillatory Element

The adaptability of the computer as designed above, to a damped oscillatory element (one with two complex poles in the left half s-plane), was found by observing the system response. To test the computer, variable damping was introduced into the oscillatory element while the undamped natural frequency (length of the radius to the complex poles in the s-plane) of the element was kept constant. The damping was varied from zero to critical. For damping between C and 0.1, the response of the undamped oscillatory element (within the accuracy of measurement). For an amplifier gain of 50, and damping between 0.2 and 1, the rise time of the response to step inputs was within ten percent of the theoretical rise time for minimum time deadbeat response. The rise time can be decreased with some increase in overshoot by decreasing the gain in the error derivative channel of the computer. See Figure 11.

Mode of Operation. The proper switch curve for a damped oscillatory element is a distorted logarithmic spiral. The semicircular switch curve gives a good response, because in the operating range it approximates the spiral. To achieve better response, the gain of the error derivative channel was decreased, changing the shape of the switch curve to two semiellipses. The semiellipse is a closer approximation to the spiral than a semicircle. The semielliptical switch curve makes the computer change sign later than the semicircular curve; therefore, the amplifier output drives the oscillatory element closer to zero error before it reverses. This causes the output of the oscillatory element to have a faster rise time, but increases its overshoot.

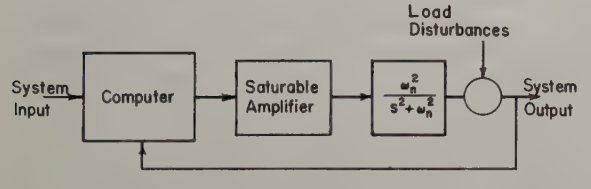


Fig. 1. Maximum effort control system.

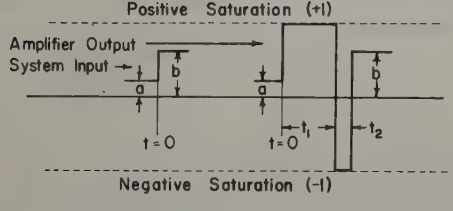


Fig. 2. Amplifier output for maximum effort control of a step input.

Conclusion

Presented is a method of design and a design for a computer to control the output of a saturating amplifier which drives an oscillatory element. The method of design uses phasors to represent the transient behavior of the step responses of the oscillatory element. Phase plane techniques are demonstrated in the development of the nonlinear computer used to control the output of the saturating amplifier.

The design presented could be used as a starting point for the design of a system with random inputs. It also provides a useful system topology for the control of an oscillatory element which cannot be damped. One possible application of this system would be in the control of phugoid low frequency oscillations in the pitch axis of an aircraft.

Acknowledgements

The author wishes to express his appreciation to Professor O. J. M. Smith for his guidance in the development of the material presented.

Bibliography

1. Smith, Otto J. M., Feedback Control Systems, New York: McGraw-Hill, 1958, Chapter 15.

2. Bushaw, D., Differential Equations with a Discontinuous Forcing Term, Stevens Institute Technological Report 469, 1952.

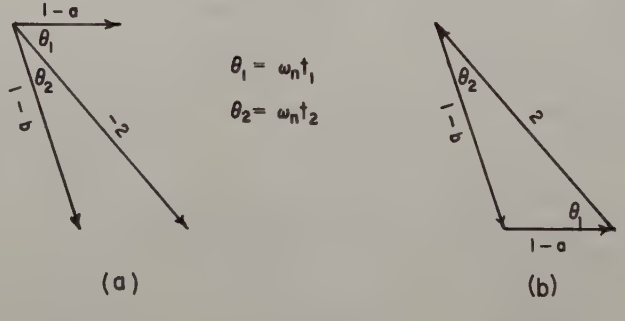


Fig. 3. (a) Phasor representation of transients introduced in the undamped element by the amplifier output shown in Fig. 2; (b) vector addition of the phasors in (a) to obtain the t_1 and t_2 necessary for deadbeat response.

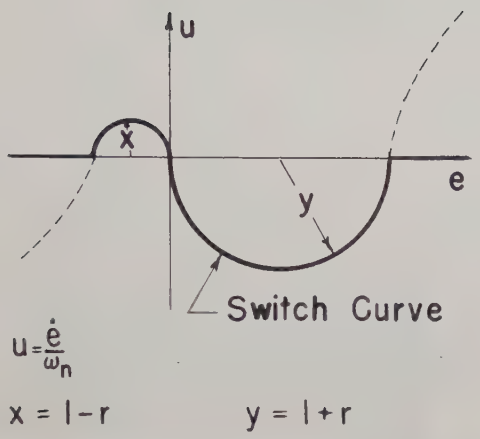


Fig. 4. Switch curve for undamped element. (r = system input)

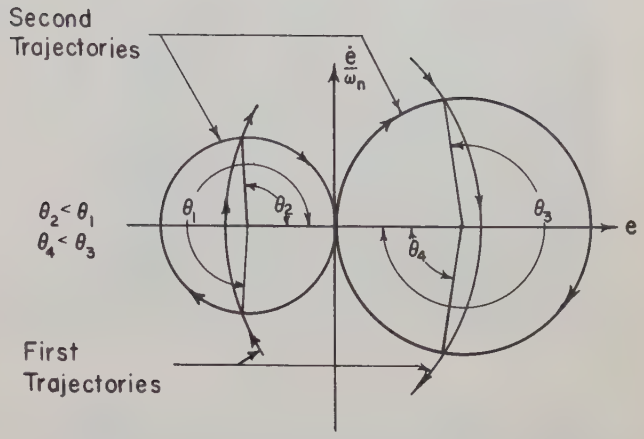


Fig. 5. Construction to find the minimum time path for a point on the first trajectory to the origin. The time spent on the second trajectories is proportional to the angles θ_1, etc. The minimum time path uses the portions of the second trajectories which are in the second and fourth quadrants.

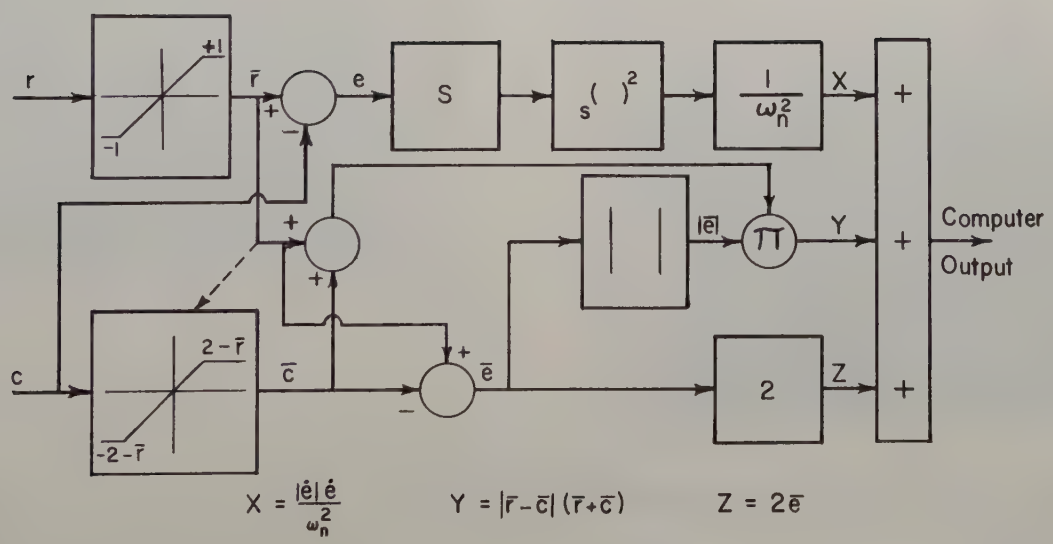


Fig. 6. Computer for maximum effort control of an undamped element (r = system input; c = system output; clipped values are represented by $\bar{r}$ and $\bar{c}$).

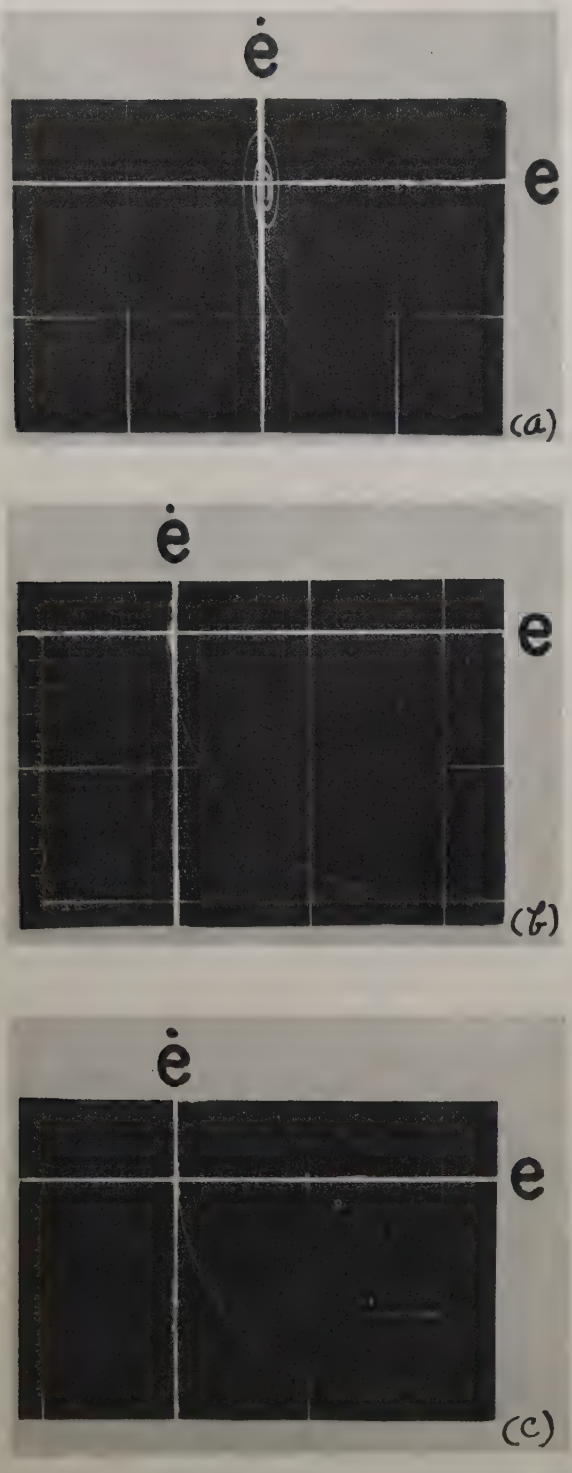


Fig. 7. Effect of amplifier gain (K) on system response to a step input (r = 0.5):
(a) K = 5;
(b) K = 50;
(c) K = 1,000.

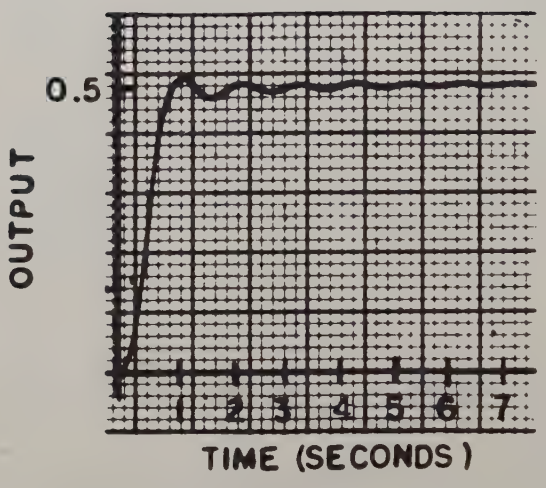


Fig. 8. Step response with gain of 50.

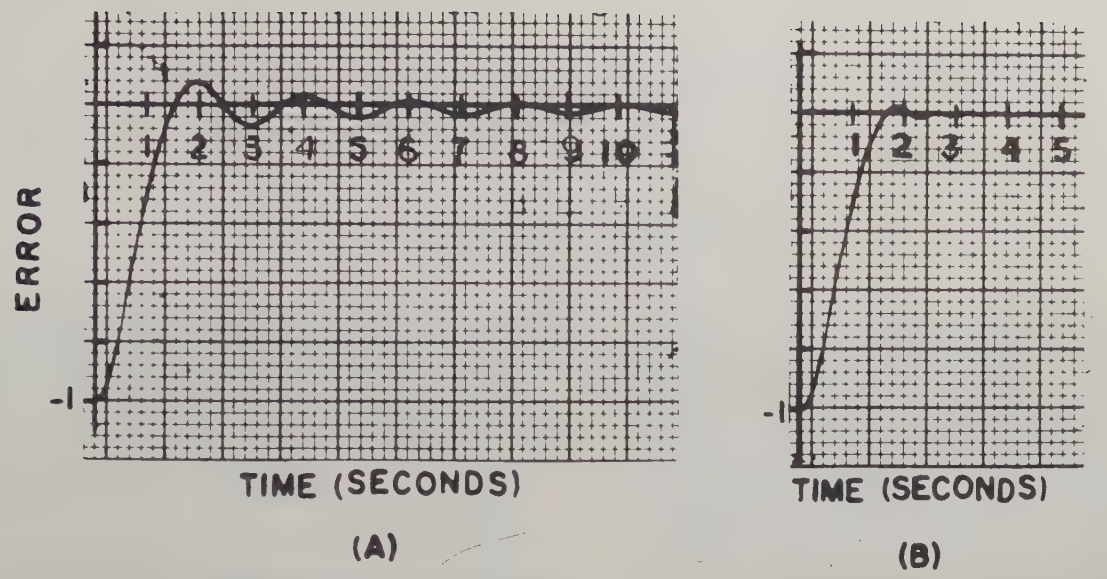


Fig. 9. Effect of amplifier gain on system response: (a) gain of 5. (b) positive feedback around amplifier acts to increase gain to more than 1,000.

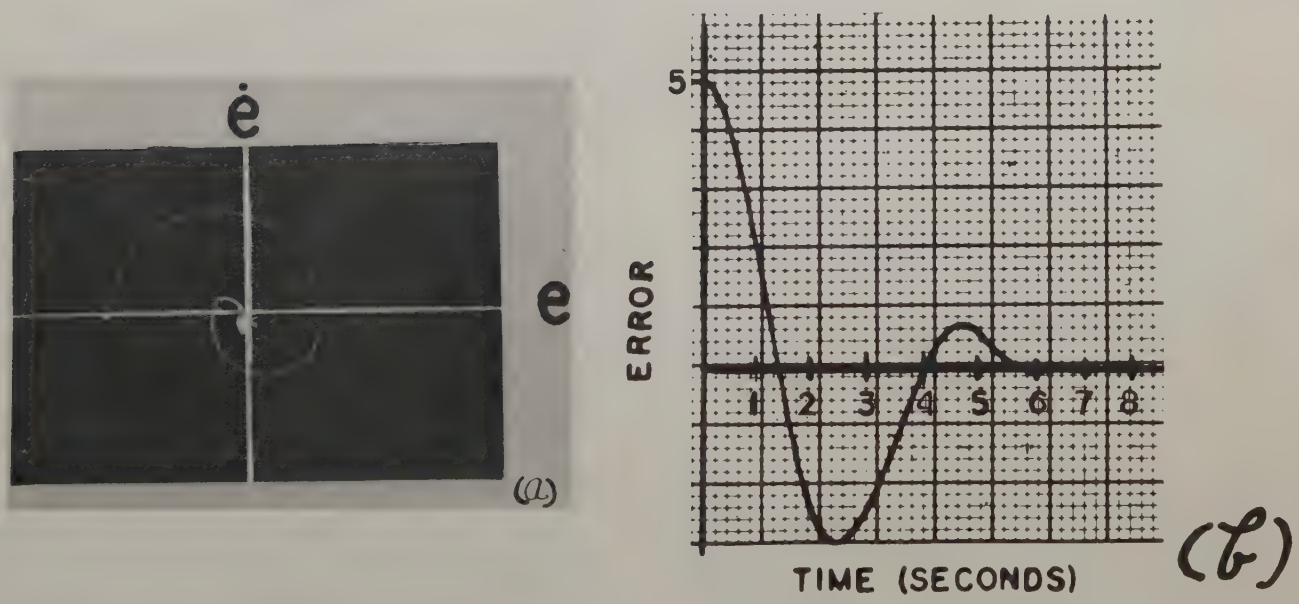


Fig. 10. System response to a load disturbance:
(a) phase plane;
(b) time response.

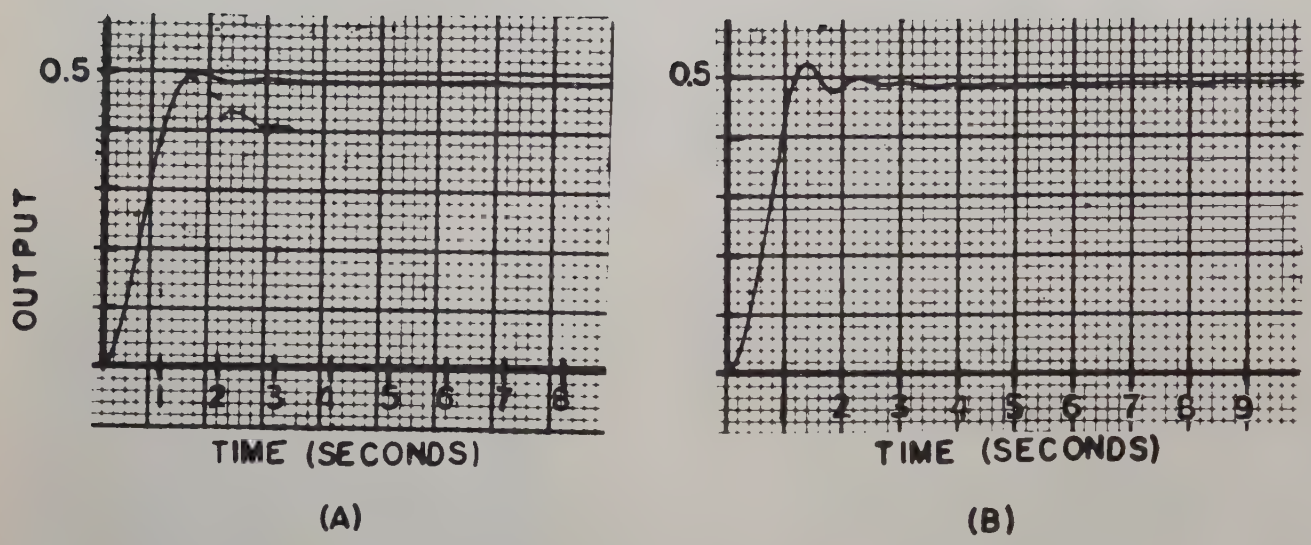


Fig. 11. Response of system containing a damped oscillatory element: (a) error derivative gain same as in control of undamped element; (b) error derivative gain is 0.4 of (a).

IDENTIFICATION AND COMMAND PROBLEMS IN ADAPTIVE SYSTEMS

by

E. Mishkin & R.A. Haddad
Microwave Research Institute
Polytechnic Institute of Brooklyn
Brooklyn 1, New York

Abstract

In order to satisfy stringent performance requirements in a dynamic process, a computer is incorporated as a central element in the feedback loop. The computer performs the dual task of identifying or measuring the process'dynamics, and thence generating an appropriate command or actuating signal so as to satisfy the overall specifications. The family of singularity functions (steps, ramps, confluent parabolas) is used as the command signal. The process' dynamics are monitored and identified by the computer without recourse to interrupting test signals such as periodic impulses or white noise. The stored energy term inherent in many measurement problems in continuous processes is accounted for in a novel manner.

I. Introduction

The term "adaptive" control system has appeared only recently in the literature on feedback control systems, and as such, is still subject to various interpretations by different authors. The primary need for adaptive systems arises from the difficulties encountered in the design of control systems for processes whose dynamic parameters vary, often in an unpredictable manner. The conventional feedback system often employs a large loop gain to swamp out the dependence of the system transmission on the varying parameters; certain model configurations have also been utilized in an attempt to nullify the effect of the varying parameters on the system performance. The design objective would then be to reduce the sensitivity of transmission with respect to some critical parameter over the expected frequency range of the input signal.

The adaptive approach would aim to maintain a prescribed sensitivity or performance criterion in the face of process changes by appropriately varying the compensating networks, or in effect, the actuating signal to the process. The adaptive viewpoint would be especially suited to the design of controllers for processes whose dynamics are not completely known in advance.

The following definition of an adaptive system has been suggested to the authors by J.G. Truxal: a feedback control system is adaptive if the sensitivity with respect to a variable x is zero over an interval in x of non-zero magnitude.

It is obvious that any realizable physical means used in reducing the system sensitivity cannot be made to react instantaneously to changes in the control process dynamics. The ideal expressed in the above definition is, therefore, only approximately realizable. The use of modern computer circuits, both analog and digital, can make this approximation quite good.

The nullification of system sensitivity with respect to any changes in the process dynamics calls for:

(1) The identification of the process, and

(2) The modification of the signal commanding the process such that the desired system performance is achieved.

The adaptive problem therefore logically splits into two halves - that of identification and command. A computer which performs this two-fold task is incorporated as a central element of the overall control system. The specific design of the computing circuits is based on the equations developed in Sections II and III of this paper.

II. The Identification Problem

This problem is treated at length in the literature by several authors (See references 1 thru 6). There is, of course, a common denominator for all attempts at formulating a mathematical model of the process' dynamics. Some properties of the excitation, and response signals must be combined together in some fashion to yield approximate numerical information regarding the process' dynamics. For a linear, time-invariant system the convolution integral

$$c(t) = \int_{-\infty}^{t} g(\tau)\, a(t-\tau)\, d\tau \qquad (1)$$

must be solved to obtain g(t), the process' impulse response if a(t), the excitation, and c(t), the response, are known.

Of course, the convolution or superposition integral is valid only for linear systems. However, if the interval of integration, T, is chosen small enough such that the process' dynamics do not vary appreciably over this interval, then the convolution integral may be extended to non-linear systems. It should be emphasized that the validity for such an assumption must be founded on some à priori information regarding the process' variations which whould set an upper limit on T. The lower limit of the interval must be set by the correlation time of the noise imbedded in the system.

For the system under consideration, the computer monitors the signals c(t) and a(t) and to some approximation solves the convolution integral for the process' unit step and/or impulse response every T seconds. The computer will then express the computed impulse response $g_c(t)$ (or step response) as a set of specific numerical values, each of which represents the solution of an equation by the computer. The accuracy of the approximation is commensurate with the type and size of the computer.

Fig. 1 represents diagramatically, an embodiment of our approach. The actuating signal a(t) is generated by the computer which also identifies the process. The problem of determining an appropriate command signal will be deferred until Section III.

If the unit step response; u(t) of the controlled process is used in lieu of the unit impulse response, then the convolution integral (1) reads

$$c(t) = \int_{-\infty}^{t} a'(\tau)\, u(t-\tau)d\tau \qquad (2)$$

or

$$c(t) = \int_{-\infty}^{0^-} a'(\tau)\, u(t-\tau)d\tau + \int_{0}^{t} a'(\tau)\, u(t-\tau)d\tau$$
$$= I_1(t) + I_2(t) \qquad (3)$$

The integral $I_1(t) = \int_{-\infty}^{0^-} a'(\tau)\, u(t-\tau)$ embodies the stored energy in the output at time t due to past inputs before t = 0. It is this term which has been particularly troublesome in the measurement problem in other proposed adaptive systems. The problem of accounting for the stored energy can be resolved in the following fashion. If $I_1(t)$ is expanded in a Taylor series about the origin ($t=0^-$) one obtains

$$I_1(t) = I_1(0) + I_1'(0)t + \frac{I_1''(0)t^2}{2!} + \cdots \qquad (4)$$

where

$$I_1(0) = \int_{-\infty}^{0^-} a'(\tau)\, u(-\tau)d\tau = c(t)\Big|_{t=0^-} = c(0^-)$$

$$I_1'(0) = \int_{-\infty}^{0^-} a'(\tau)u'(-\tau)d\tau = \frac{dc(t)}{dt}\Big|_{t=0^-} = c'(0^-) \qquad (5)$$

$$I_1''(0) = \int_{-\infty}^{0^-} a'(\tau)u''(-\tau)d\tau = \frac{d^2c(t)}{dt^2}\Big|_{t=0^-} = c''(0^-)$$

Hence

$$I_1(t) = c(0^-) + tc'(0^-) + \frac{t^2c''(0^-)}{2!} + \cdots$$
$$= \sum_{i=0}^{\infty} t^i \frac{c^i(0^-)}{i!} \qquad (6)$$

The operation dictated by Eq. (6) involves the sum of an infinite Taylor series. In practice, only a few terms need be taken since the series can be made to converge rapidly in the interval $0 < t < T$ by a judicious choice of T. Henceforth, n will be used as the upper limit on the summation index. Eq. (3) becomes

$$c(t) = \sum_{i=0}^{n} \frac{t^i c^i(0^-)}{i!} + I_2(t) \qquad (7)$$

The analytic evaluation of $I_2(t)$ may be simplified if the command signal a(t) is synthesized from elementary units, say the family of singularity functions. The following forms for a(t) will be considered:

(1) A set of impulses of varying areas spaced T seconds apart.

(2) Successive step-functions, realized by passing a string of pulses through a zero-order holding device. a(t) is then a "staircase" function.

(3) A piecewise-linear time function obtained by generating ramp functions every T seconds.

(4) A series of confluent parabolas.

(5) A combination of such singularity functions.

(a) "Staircase" Command Signal

For the case at hand, let a(t) be the staircase function shown in Fig. 2a. Then a'(t) is a string of impulses occurring at t = 0, T, 2T..., etc. as shown in Fig. 2b. In the interval $0^- \; t \; T^-$, $I_2(t)$ can now be readily evaluated to obtain

$$I_2(t) = \int_{0^-}^{t} (a_0 - a_{-1})\delta(\tau)u(t-\tau)d\tau = \Delta a_0 u(t)_1 \,, \quad 0 < t < T \qquad (9)$$

where

$$\Delta a_0 = (a_0 - a_{-1}) \qquad (10)$$

$u(t)_1$ is the process' unit step response in the interval $0 < t < T$. If we let $t = T^-$ in Eqs. (6) and (9), Eq. (7) becomes

$$c(T^-) = \sum_{i=o}^{n} \frac{T^i c^i(0^-)}{i!} + \Delta a_0\, u(T)_1 \qquad (11)$$

From which

$$u(T)_1 = \frac{c(T^-) - \sum_{i=o}^{n} \frac{T^i c^i(0^-)}{i!}}{\Delta a_0} \qquad (12)$$

$u(T)_1$ is interpreted as the response of the process to a unit step applied T seconds earlier. The subscript 1 implies that the computation is performed in the first measurement interval.

Similarly,

$$c(2T^-) = \sum_{i=o}^{n} \frac{T^i c^i(T^-)}{i!} + \Delta a_1\, u(T)_2$$

$$u(T)_2 = \frac{c(2T^-) - \sum_{i=o}^{n} \frac{T^i c^i(T^-)}{i!}}{\Delta a_1} \qquad (13)$$

$$\vdots$$

$$u(T)_k = \frac{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i\left[(k-1)T^-\right]}{i!}}{\Delta a_{k-1}}$$

Thus Eq. (13) identifies the process step response at the k'th measuring period in terms of the value of the output at kT^-, the values of the output and its derivatives at the preceding $\left[(k-1)T\right]$ sampling instant, and the values of the command signal in the two preceding sampling intervals.

A computer will now be called upon to carry out the dictates of Eq. (13) and yield the identifying function $u(T)_k$. The computer selected or designed must carry out the computation in a time duration much lessthan the sampling interval T.

(b) Piecewise-Linear Command Signal

Let the command signal be the piecewise-linear curve of Fig. 3a; m_k is the slope of the straight line segment for $kT < t < (k+1)T$. Fig. 3b illustrates the derivative a'(t). By an analysis similar to that of Section (a), it is clear that

$$c(t) = \sum_{i=o}^{n} \frac{t^i c^i(0^-)}{i!} + m_0 \int_{o^-}^{t} u(t-\tau)d\tau \quad 0 < t < T \qquad (14)$$

where

$$\int_o^t u(t-\tau)d\tau = \int_o^t u(\tau)d\tau = u^{-1}(t)_1 \qquad (15)$$

is the response of the process to a unit ramp input for $0 < t < T$. The subscript has the same significance as that for the unit step response considered previously. In the k'th interval, $(k-1)T < t < kT$

$$c(t) = \sum_{i=o}^{n} \frac{\underline{t}^i c^i\left[(k-1)T^-\right]}{i!} + m_{k-1} \int_{(k-1)T}^{t} u(t-\tau)d\tau \qquad (16)$$

where

$$\underline{t} = t - (k-1)T \qquad (17)$$

and

$$u^{-1}(\underline{t})_k = \int_{(k-1)T}^{t} u(t-\tau)d\tau = \int_o^{\underline{t}} u(\tau)d\tau \qquad (18)$$

is the process' unit ramp response in the k'th interval. Solving Eq. (16) for the ramp response at $t = kT^-$, (or $\underline{t} = T^-$) one obtains the identifying function

$$u^{-1}(T)_k = \frac{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i\left[(k-1)T^-\right]}{i!}}{m_{k-1}} \qquad (19)$$

(c) Confluent Parabolas as Command Signals

Let a(t) consist of a series of confluent parabolas as shown in Fig. 4 where

$$a(t) = a_{k-1} + \tfrac{1}{2} m_{k-1} \left[t-(k-1)T\right]^2, \text{ for } (k-1)T < t < kT \qquad (20)$$

and

$$a_k = a_{k-1} + \tfrac{1}{2} m_{k-1} T^2 \qquad (21)$$

$$u^{-2}(\underline{t}) = \int_o^{t-(k-1)T} u^{-1}(\tau)d\tau = \int_o^{\underline{t}} d\tau \int_o^{\tau} u(x)dx \qquad (22)$$

at $t = kT^-$, the process' parabolic response can be shown to be

$$u^{-2}(T)_k = \frac{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i\left[(k-1)T^-\right]}{i!}}{m_{k-1}} \qquad (23)$$

(d) Combination of Steps and Ramps as Command Signals

The identification problem is somewhat complicated when two singularity functions are synthesized as the command signal. However, as will be shown in Section III, such a combination permits an additional degree of freedom in satisfying the system specifications. Let the command signal consist of steps and ramps as shown in Fig. 5a. The derivative a'(t) shown in Fig. 5b then consists of impulses and steps.

In the interval, $(k-1)T^- < t < kT^-$,

$$a'(t) = \left[a_{k-1}-(a_{k-2}+m_{k-2}T)\right]\delta\left[t-(k-1)T\right] + m_{k-1} \quad (24)$$

and the output during this interval $\left[\underline{t} = t-(k-1)T\right]$ is

$$c(t) = \sum_{i=o}^{n} \frac{\underline{t}^i c^i\left[(k-1)T^-\right]}{i!} + \int_{(k-1)T^-}^{t} a'(t)\, u(t-\tau)\,d\tau$$

$$= \sum_{i=o}^{n} \frac{\underline{t}^i c^i\left[(k-1)T^-\right]}{i!} \quad (25)$$

$$+ \left[a_{k-1} - (a_{k-2}+m_{k-2}T)\right]u(\underline{t}) + m_{k-1}\, u^{-1}(\underline{t})$$

In anticipation of the specifications to be set down in Section III, the derivative of the output is also determined; for convenience, let the derivative be designated by b(t), i.e., b(t) = dc/dt. Then

$$b(t) = \sum_{i=o}^{n} \frac{\underline{t}^i b^i\left[(k-1)T^-\right]}{i!}$$

$$+ \left[a_{k-1}-(a_{k-2}+m_{k-2}T)\right] g(\underline{t})_k + m_{k-1}\, u(\underline{t})_k \quad (26)$$

At $t = kT^-$, Eqs. (25) and (26) become

$$c(kT^-) = \sum_{i=o}^{n} \frac{T^i c^i\left[(k-1)T^-\right]}{i!} \quad (27a)$$

$$+ \left[a_{k-1}-(a_{k-2}+m_{k-2}T)\right]u(T)_k + m_{k-1}u^{-1}(T)_k$$

$$b(kT^-) = \sum_{i=o}^{n} \frac{T^i c^i\left[(k-1)t^-\right]}{i!} \quad (27b)$$

$$+ \left[a_{k-1}-(a_{k-2}+m_{k-2}T)\right] g(T)_k + m_{k-1}u(T)_k$$

where

$g(T)_k$ = response to a unit impulse applied at $t = (k-1)T$ and evaluated at $t = kT$

$u(T)_k$ = response to a unit step

$u^{-1}(T)_k$ = response to a unit ramp

The identifying quantities, $g(T)_k$, $u(T)_k$, and $u^{-1}(T)$ are given in Eq. (27) by two simultaneous equations. However, if $g(T)_k$, $u(T)_k$, and $u^{-1}(T)$ are considered as discrete variables, then, of course, the two equations in three unknowns cannot yield unique solutions. This dilemma may be resolved by recalling that Eqs. (27) were obtained from two integro-differential equations which were evaluated at discrete intervals in time. Furthermore, $u^{-1}(t)$ is recognized as the integral of u(t) which, in turn, is the integral of g(t). The two simultaneous equations can be simulated by analog techniques to yield u(t), etc. on a continuous basis. Readouts taken every kT seconds will yield $g(T)_k$, $u(T)_k$ and $u^{-1}(T)_k$.

Consider Eqs. (25) and (26) and define

$$\sigma_{k-1} = \left[a_{k-1}-(a_{k-2}+m_{k-2}T)\right] \quad (28)$$

$$\gamma_{k-1}(\underline{t}) = \sum_{i=o}^{n} \frac{\underline{t}^i c^i\left[(k-1)T^-\right]}{i!}$$

and then

$$\gamma'_{k-1}(\underline{t}) = \sum_{i=o}^{n} \frac{\underline{t}^i b^i\left[(k-1)T\right]}{i!}$$

Substituting the definitions of Eqs. (28) into Eqs. (25) and (26), we obtain

$$c(t) - \gamma_{k-1}(\underline{t}) = \sigma_{k-1}u(\underline{t}) + m_{k-1}\, u^{-1}(\underline{t}) = \alpha(t) \quad (29)$$

$$b(t) - \gamma'_{k-1}(\underline{t}) = \sigma_{k-1}g(\underline{t}) + m_{k-1}\, u(\underline{t}) = \alpha'(t)$$

Eqs. (28) readily lend themselves to analog simulation and hence a continuous solution for $u(\underline{t})$, etc. can be obtained. The output $c(t)$ and its derivative are continuously monitored; inspection of Eqs. (28) reveals that σ_{k-1} is known in terms of the past input, while $\gamma_{k-1}(\underline{t})$ can be generated with a knowledge of $c(t)$ and its derivatives at $t = (k-1)T$ (or $\underline{t} = 0$).

The operation of this analog system will be continuous except for readouts taken at every $t = kT$ seconds which will then yield the required values of $u(T)_k$, $g(T)_k$, and $u^{-1}(T)_k$. Eqs. (29) can combined to yield

$$-\sigma_{k-1}^2 \frac{du}{dt} = m_{k-1}\, c(t) - m_{k-1}^2 \int u(\tau)d\tau - \sigma_{k-1} c'(t) \tag{30}$$

A possible scheme for simulating this equation in block diagram form is shown in Fig. 6.

III. The Command Problem

Once the computer has solved the identification problem, then it will be called upon to generate an appropriate command signal to drive the process. Since it has been decided to use the family of singularity functions as the form for the command signal, then on the basis of the specifications, the computer is to decide, and generate, the required steps, ramps, etc. after each sampling interval.

Let the desired output of the controlled process be $c_d(t)$ where $c_d(t)$ may be some linear or nonlinear function of the input $r(t)$. This desired output, $c_d(t)$, may be generated by applying the input $r(t)$ to an appropriate model representing the desired operation. This desired output can then be monitored continuously or periodically, and can be operated upon in any desired fashion.

Various control strategies can be worked out. We shall limit our considerations at present to two sets of specifications. These are

(a) Let the output $c(t)$ approximately follow the desired output $c_d(t)$, at the sampling instants, but with a unit-time delay T, i.e.,

$$c(kT) = c_d[(k-1)T] \tag{31}$$

(b) In this instance, the computer is required to match the output signal and its slope to the desired output and slope at the sampling instants with a unit time delay; i.e.,

$$\begin{aligned} c(kT) &= c_d[(k-1)T] \\ c'(kT) &= c_d'[(k-1)T] \end{aligned} \tag{32}$$

A staircase command signal is generated by the computer to satisfy the specifications of Eq. (31), while a combination of steps and ramps are required for Eqs. (32). Of course, the specifications of Eqs. (32) impose a greater degree of complexity on the computer than those of Eq. (31).

(a) Implementation of Eq. (31) with a Staircase Command Signal

Let the command signal be the staircase function of Fig. 2. Eqs. (11) and (12) are rewritten here for immediate reference.

$$c(T^-) = \sum_{i=o}^{n} \frac{T^i c^i(0^-)}{i!} + \Delta a_o\, u(T)_1 \tag{11}$$

$$u(T)_1 = \frac{c(T^-) - \sum_{i=o}^{n} \frac{T^i c^i(0^-)}{i!}}{\Delta a_o} \tag{12}$$

If the strategy of Eq. (31) is to be satisfied, then $c(2T^-) = c_d(T)$; substitution of Eq. (31) into Eq. (11) yields

$$c_d(T) = c(2T^-) = \sum_{i=o}^{n} \frac{T^i c^i(T^-)}{i!} + \Delta a_1\, u(T)_2 \tag{33}$$

Eq. (32) is to be solved for Δa_1, the height of the step to be applied at $t = T$. But $u(T)_2$ is the response of the process at $t = 2T$ to a unit step applied at $t = T$, and is an unknown in Eq. (32). This dilemma can be resolved by assuming that $u(T)_2$ is equal to $u(T)_1$, the previously measured step response. Such an assumption is valid provided that the process' dynamics do not vary appreciably over one sampling interval. Hence let

$$u(T)_2 \doteq u(T)_1 \tag{34}$$

Combining Eqs. (34), (33) and (12) gives the required height of the step command at $t = T$ or

$$\Delta a_1 = \frac{c_d(T) - \sum_{i=o}^{n} \frac{T^i c^i(T^-)}{i!}}{c(T^-) - \sum_{i=o}^{n} \frac{T^i c^i(0^-)}{i!}}\, \Delta a_o \tag{35}$$

The general term is

$$\Delta a_k = \frac{c_d(kT) - \sum_{i=o}^{n} \frac{T^i c^i[kT^-]}{i!}}{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i[(k-1)T^-]}{i!}} \Delta a_{k-1} \qquad (36)$$

where n = number of terms in the Taylor expansion, Δa_{k-1} is the height of the step that was applied at $t = (k-1)T$ and Δa_k is the height of the step demanded at $t = kT$. The computer is to solve Eq. (36) at each $t = kT - 0$ for Δa_k and thence to generate Δa_k to command the process.

(b) Implementation of Eq. (31) with Piecewise Linear Command Signal

For reasons set down in (a) it is assumed that

$$u^{-1}(T)_k \doteq u^{-1}(T)_{k-1} \qquad (37)$$

Combining Eqs. (31), (16), (19) and (37) in a fashion similar to that of Section II a, yields m_k, the required slope of the ramp command signal in the interval $kT < t < (k+1)T$.

$$m_k = \frac{c_d(kT) - \sum_{i=o}^{n} \frac{T^i c^i(kT^-)}{i!}}{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i[(k-1)T^-]}{i!}} m_{k-1} \qquad (38)$$

(c) Confluent Parabola Command Signal for Eq. (31)

The recurrence equation obtained by following the procedure outlined previously to satisfy the strategy of Eq. (31) is therefore,

$$m_k = \frac{c_d(kT) - \sum_{i=o}^{n} \frac{T^i c^i(kT^-)}{i!}}{c(kT^-) - \sum_{i=o}^{n} \frac{T^i c^i[(k-1)T^-]}{i!}} m_{k-1} \qquad (39)$$

(d) Implementation of the Control Strategy of (Eq. 32) by Using Combinations of Steps and Ramps as a Command Signal

The specifications of Eq. (32) require an additional degree of freedom in the command signal. These can be met by le tting the command consist of a sum of any two singularity functions, as for example the combination of staircase and piecewise-linear segments shown in Fig. 5 and analyzed in Section I d. At this point it is assumed that the identification problem has been solved so that $g^{-1}(T)_k$, $u(T)_k$, and $u^{-1}(T)$ are known. The computer must now determine the slope of the ramp m_k, and the height of the step a_k required at $t = kT$ to satisfy the specifications. Combining Eq. (32) with (27) gives

$$c[(k+1)T^-] = \sum_{i=o}^{n} \frac{T^i c^i(kT^-)}{i!} +$$

$$[a_k - (a_{k-1} + m_{k-1}T)]\, u(T)_{k+1} + m_k\, u^{-1}(T)_{k+1} = c_d(kT) \qquad (40)$$

$$b[(k+1)T^-] = \sum_{i=o}^{n} \frac{T^i b^i(kT^-)}{i!} +$$

$$[a_k - (a_{k-1} + m_{k-1}T)]\, g(T)_{k+1} + m_k\, u(T)_{k+1} = b_d(kT)$$

If it is assumed that

$$u(T)_{k+1} \doteq u(T)_k$$

$$g(T)_{k+1} \doteq g(T)_k \qquad (41)$$

$$u^{-1}(T)_{k+1} \doteq u^{-1}(T)_k$$

then Eqs. (40) can be solved for a_k and m_k.

Let us define

$$\gamma_k(T) = \sum_{i=o}^{n} \frac{T^i c^i(kT^-)}{i!}$$

$$\gamma'_k(T) = \sum_{i=o}^{n} \frac{T^i b^i(kT^-)}{i!} \qquad (42)$$

$$\rho_{k-1} = (a_{k-1} + m_{k-1}T)$$

Furthermore, we define

$$c_k(kT) - \gamma_k(T) + \rho_{k-1}\, u(T)_k = \xi_k$$

$$b_d(kT) - \gamma'_k(T) + \rho_{k-1}\, g(T)_k = \eta_k \qquad (43)$$

so that Eqs. (40) take the form

$$\begin{bmatrix} u(T)_k & u^{-1}(T)_k \\ g(T)_k & u(T)_k \end{bmatrix} \begin{bmatrix} a_k \\ m_k \end{bmatrix} = \begin{bmatrix} \xi_k \\ \eta_k \end{bmatrix} \qquad (44)$$

whose solution is

$$\begin{bmatrix} a_k \\ m_k \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} u(T)_k & -u^{-1}(T)_k \\ -g(T)_k & u(T)_k \end{bmatrix} \begin{bmatrix} \xi_k \\ \eta_k \end{bmatrix} \qquad (45)$$

and

$$\Delta = \begin{vmatrix} u(T)_k & u^{-1}(T)_k \\ g(T)_k & u(T)_k \end{vmatrix} = u(T)_k^2 - g(T)_k\, u^{-1}(T)_k \qquad (46)$$

A computer must now be programmed to solve Eqs. (45) for a_k and m_k, and thence to generate these at $t = kT$ as the command signal to the process.

IV. Computer Circuitry for Instrumentation of Eq. (36)

It was decided to satisfy the specifications of Eq. (31) through the use of the staircase function discussed in Sections II(a) and III(a). An hybrid computer was designed to solve Eq. (36) for the case when the Taylor series is truncated so that only the first derivative is included. Then the equation takes the form

$$\Delta a_k = \frac{c_d(kT) - c(kT) - Tc'(kT)}{c(kT) - c[(k-1)T] - Tc'[(k-1)T]}\, \Delta a_{k-1} \qquad (47)$$

The circuitry described here to instrument Eq. (47) was designed by the late Professor M.V. Joyce. The block diagram for the computer is shown in Fig. 7 and the circuitry for each of the individual blocks is shown in the succeeding diagrams. The H block represents a holding circuit, the T and δ blocks are delay circuits; the -1 block is a simple sign-changer, the $-\sum$ is an operation amplifier used as a summer and the $\div$ and x are dividers and multipliers respectively. The signals a,b,c,d represent pulse trains to read in and read out from the various hold and delay circuits. The hold circuit shown in Fig. 8 is pulsed by the readout pulse b every T = 100 milliseconds. The time delay circuit consists of two hold circuits in cascade; the delay depends on the time separation between the read in and read out pulses. The pulse train is shown in Fig. 11(a), the pulse generator is shown schematically in Fig. 11(b).

The division operation is carried out implicity as shown in Fig. 10.

To complete the picture, the actual input r(t) would drive a model whose output would be $c_d(t)$. The model would embody the desired system characteristics, and may be built into the computer itself.

Experimental work has been temporarily stopped due to the sudden death of Professor M.V. Joyce, but it is hoped that some experimental data would be available in the near future.

V. Conclusions

Ideally, the adaptive system which has been described in this paper, will satisfy the performance specifications despite unpredictable variations in the process' dynamics. In effect, the process is forced to behave in a desired fashion through the control exerted by the command signal. To effect this arbitrary control, it has been necessary to use computing circuits whose complexity increases with the stringency of the requirements. If the adaptive approach is to be utilized, then there is a threshold level of computer facilities and correspondingly a threshold level of expense.

Obviously, one would not resort to the adaptive viewpoint to control a relatively simple process, as say a two-phase motor. The improvement in system performance is purchased at a dear price, which must be commensurate with the process to be controlled. Thus, there is a certain threshold boundary for the complexity of the controlled process and stringency of the performance specifications beneath which computer-controlled adaptive systems will not be feasible.

References

1. L. Braun, E. Mishkin and J.G. Truxal, "On the Approximate Identification of System Dynamics" (to be published)
2. R. Staffin, "Executive-Controlled Adaptive Systems", Doctoral Dissertation, Polytechnic Institute of Brooklyn.
3. L. Braun, "On Adaptive Control Systems", Doctoral Dissertation, Polytechnic Institute of Brooklyn.
4. R.E. Kalman, "Design of a Self-Optimizing Control System", Trans. of ASME, Feb. 1958.
5. H. Wallman, "Electronic Integral Transform Computer and the Practical Solution of Integral Equations", Journal of the Franklin Institute, Vol. 250, p. 45, July 1950.
6. J.A. Aseltine, et al, "A Self-Adjusting System for Optimum Dynamic Performance", 1958 IRE National Convention Record, Part 4.

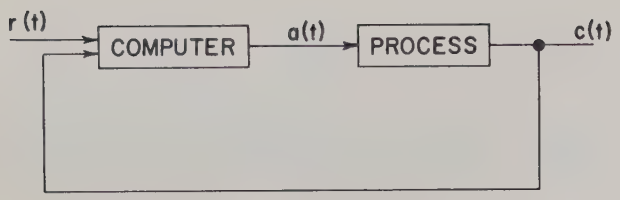


Fig. 1. Computer controlled adaptive system.

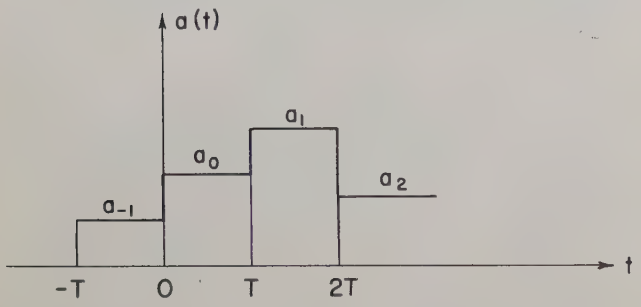


Fig. 2a. Staircase form for a(t).

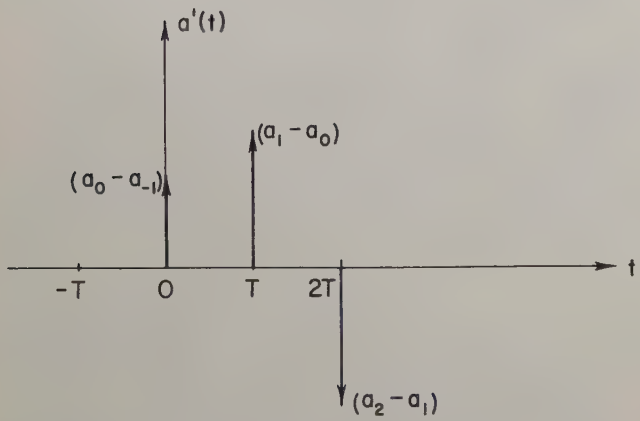


Fig. 2b. Derivative of the staircase function—a train of impulses.

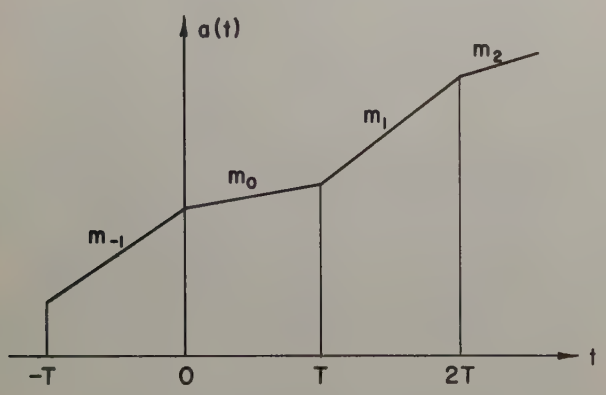


Fig. 3a. Piecewise linear form for a(t).

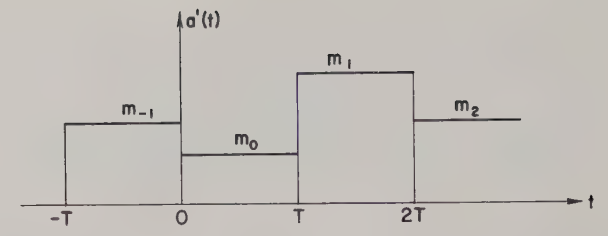


Fig. 3b. Derivative of a(t) yielding a staircase function.

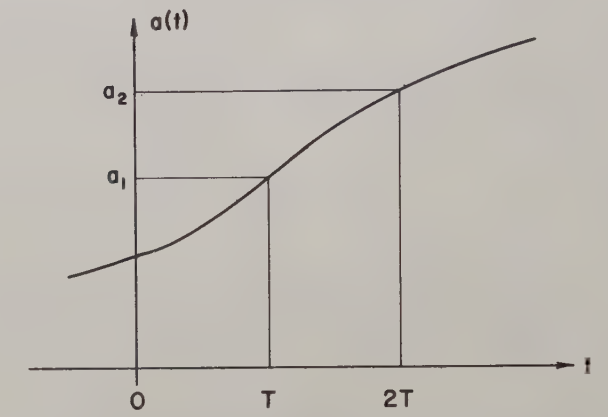


Fig. 4. Synthesis of a(t) by means of confluent parabola segments.

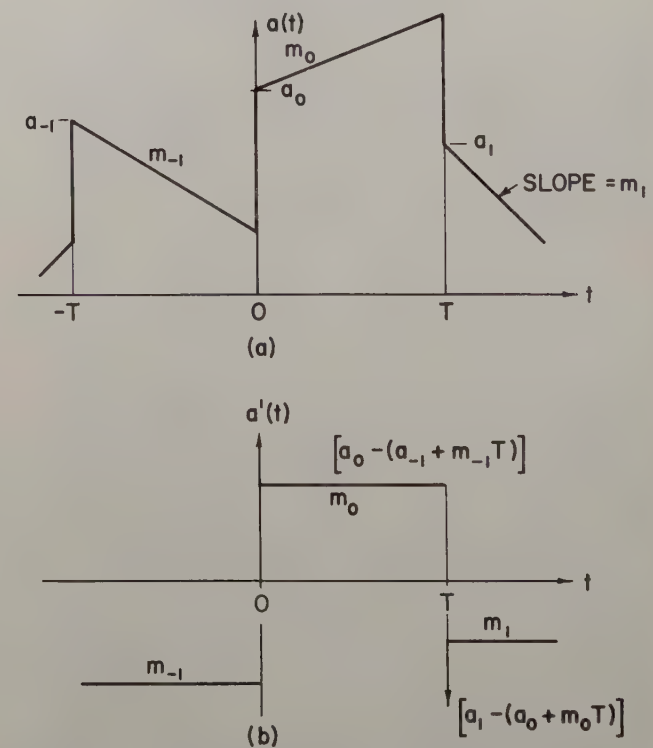


Fig. 5. Actuating signal a(t) and its derivative.

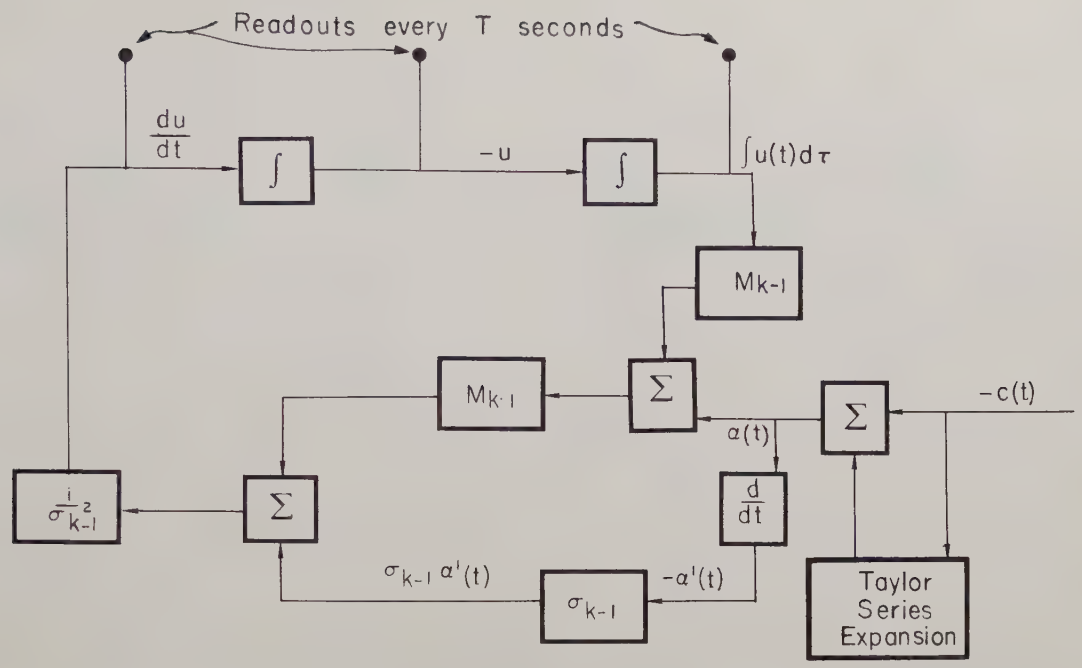


Fig. 6. Analog simulation of Eq. (30); the computer feeds potentiometers controlling m_{k-1}, and σ_{k-1}.

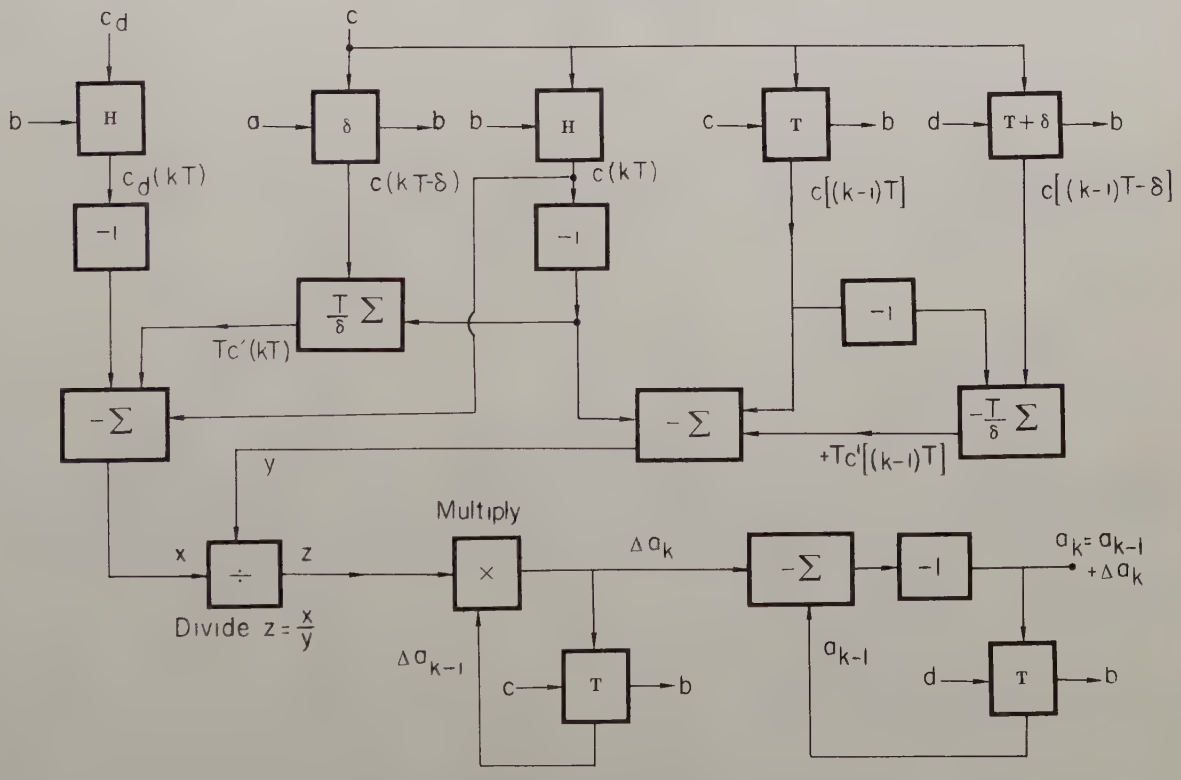


Fig. 7. Block diagram for computer instrumentation of Eq. (47).

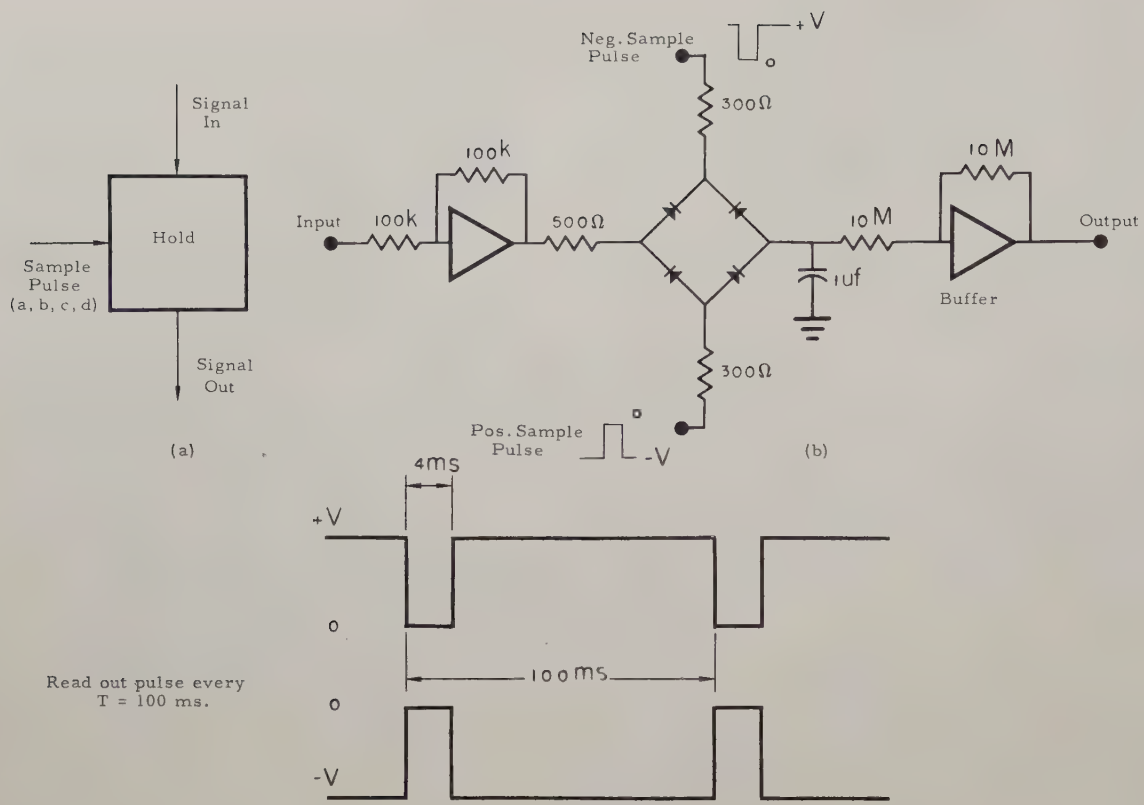


Fig. 8. Hold circuit.

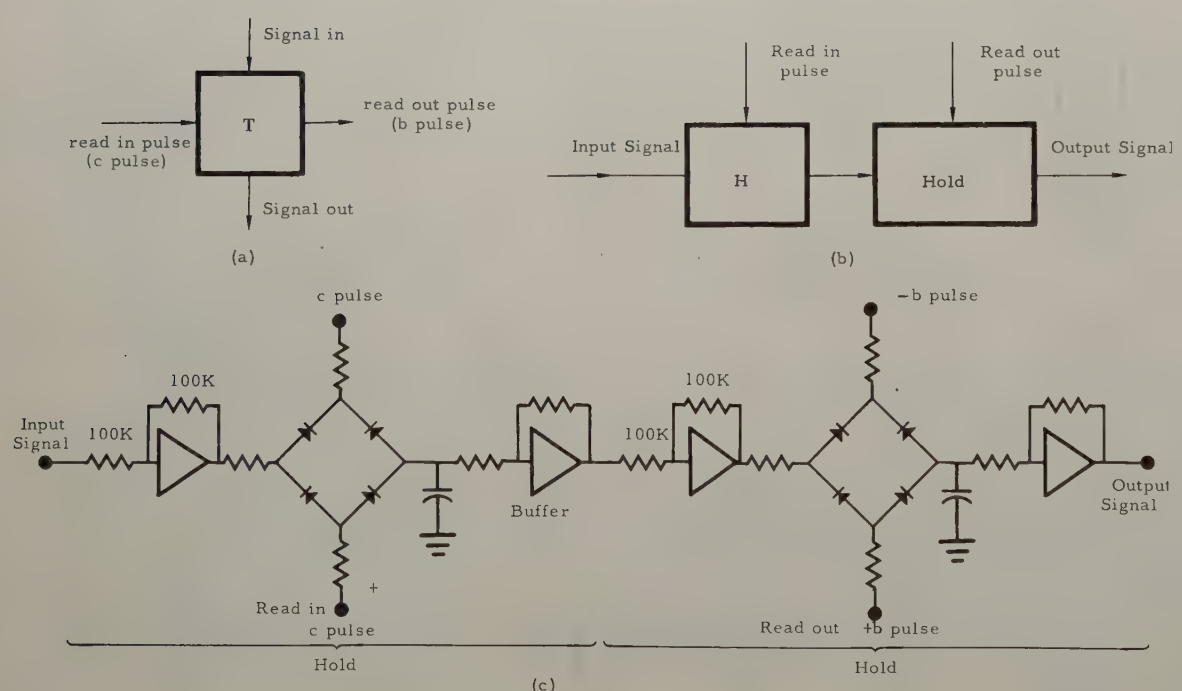


Fig. 9. Delay circuit.

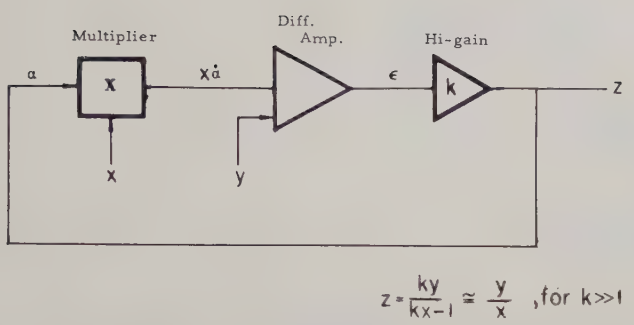


Fig. 10. Divider.

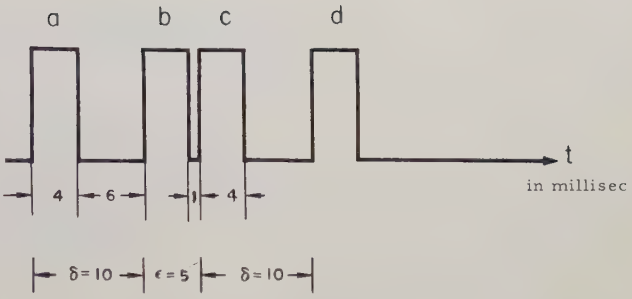


Fig. 11a. Read-in and read-out pulse train for T = 100 msec.

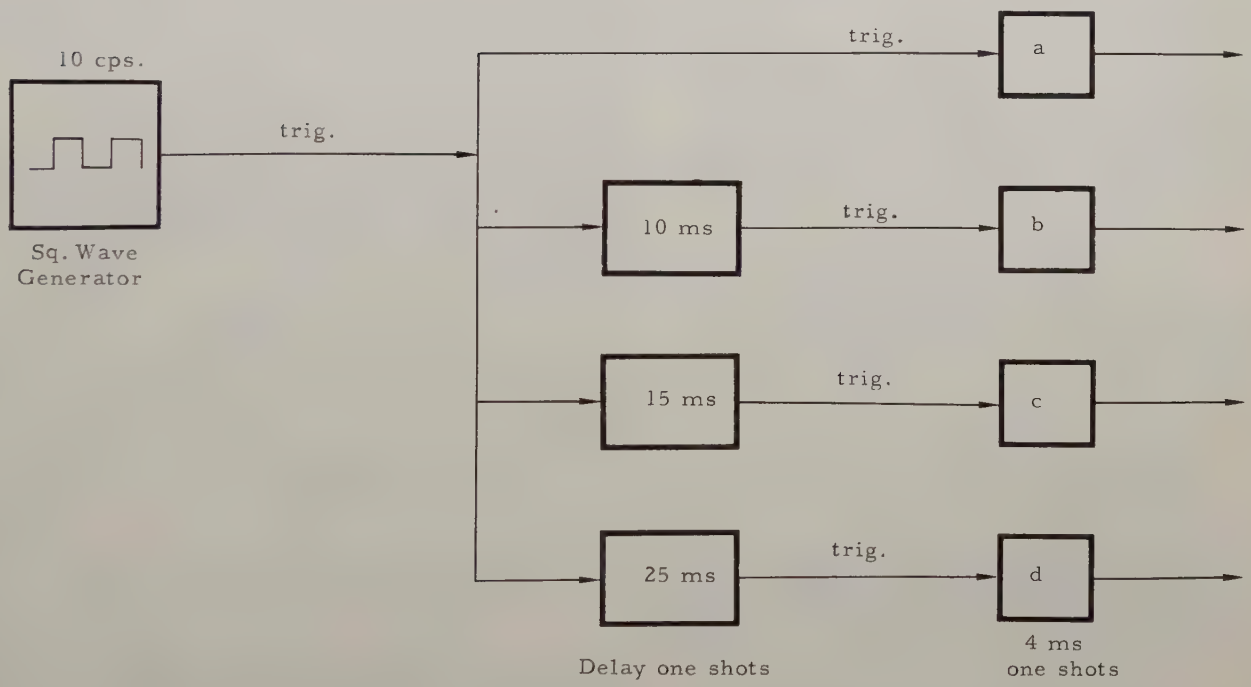


Fig. 11b. Pulse train generators.

EVALUATING RESIDUES AND COEFFICIENTS OF HIGH ORDER POLES

Dov Hazony
Case Institute of Technology
Cleveland, Ohio
formerly with Electro-Measurements, Inc.

Jack Riley
Electro-Measurements, Inc.
Portland, Oregon

Summary

The powerful Laplace transform method for transient analysis by the partial fraction expansion technique has become quite popular in the fields of circuit and servo design. The theory of residues is usually used to find the coefficients of these fractions. The process is quite simple until second and higher order poles are included in the denominator. Previously, this has required a return to the calculus to find the additional coefficients required.

This paper discloses a simple technique for finding these additional coefficients by algebraic processes. As a result both manual and machine computation can be performed more easily. The technique is described and its mathematical basis is rigorously proven.

I Introduction

This paper describes a simplified technique for the evaluation of the coefficients of partial fraction expansions. The method is based on the following relationships.

Rule: For a normalized ratio of polynomials:

A. If the denominator is one degree higher than the numerator, the sum of the residues is one.

B. If the denominator is two or more degrees higher than the numerator, the sum of the residues is zero.

The proof for these rules is given in the appendix.

Several additional rules are derived from these to permit algebraic evaluation of partial fraction coefficients in cases for which the degree of the numerator is equal or higher than that of the denominator.

The application of these rules provides a technique for finding one residue as a simple function of the others. By this means it is possible to find all of the coefficients of a partial fraction expansion by algebra even when higher order poles are involved. This technique reduces the effort of computation considerably whether the coefficients are being found directly, graphically with computational aids such as the Spirule or Nomoto's log-s plane charts, or with computers, either digital or algebraic such as the ESIAC. Before exploring this technique it will be useful to define the terms residue and normalized ratio as they are used in this paper.

Residue refers only to the coefficients of terms with first degree denominators in a partial fraction expansion. Constants, coefficients of positive powers of s, and coefficients of terms with higher degree denominators are referred to as coefficients only.

Normalization refers to the process of obtaining a coefficient of one for the highest power term in the polynomials of the numerator and denominator. This can always be accomplished by factoring out the ratio of the coefficients of the highest order terms before making the partial fraction expansion. This ratio is then used as a multiplier for all of the resulting coefficients.

II Applications of the Rule

The usefulness of this relationship of the sum of the residues for practical partial fraction expansions can be easily demonstrated by several examples.

A. First Order Poles Only:

$$\frac{(s+a)}{(s+b_1)(s+b_2)(s+b_3)} = \frac{A}{(s+b_1)} + \frac{B}{(s+b_2)} + \frac{C}{(s+b_3)}$$

The residues A and B are found by the usual technique of multiplying by $(s+b_i)$ and setting $s=-b_i$.

$$A = \frac{(a-b_1)}{(b_2-b_1)(b_3-b_1)} \qquad B = \frac{(a-b_2)}{(b_1-b_2)(b_3-b_2)}$$

then since $A+B+C=0$ the remaining residue C becomes

$$C = -(A+B)$$

Usually A and B are in numeric form so that C can be easily found. If there are only two poles and no zeros the residues will be equal in magnitude and opposite in sign. If there is only one less zero than pole the sum of the normalized residues will be one. If the degree of the numerator is equal to or higher than that of the denominator algebraic methods for finding the coefficients are still possible. Some of the techniques which can be used are given in section III.

B. One Second Order Pole:

$$\frac{1}{(s+b_1)^2(s+b_2)} = \frac{A}{(s+b_1)^2} + \frac{B}{(s+b_1)} + \frac{C}{(s+b_2)}$$

A and C can be found as above.

$$A = \frac{1}{(b_2-b_1)} \qquad C = \frac{1}{(b_1-b_2)^2}$$

By our definition only B and C are residues. Since $B+C = 0$:

$$B = -C$$

Thus an answer is obtained quickly and simply by inspection instead of having to make a complicated differentiation.

The evaluation of coefficients in the presence of one second order pole can be generalized. Our problem can be stated as follows:

Given:

$$\frac{\prod_i (s+a_i)}{(s+b_0)^2 \prod_j (s+b_j)} = \frac{C}{(s+b_0)^2} + \frac{R_0}{(s+b_0)} + \sum_1^n \frac{R_k}{(s+b_k)}$$

$i = 1, 2, \cdots, p$
$j = 1, 2, \cdots, n$
$p < (n+2)$

Find: C, R_0, all R_k.

The values of R_k can be found as before and can be written as:

$$R_k = \left.\frac{(s+b_k)\prod_i (s+a_i)}{(s+b_0)^2 \prod_j (s+b_j)}\right|_{s=-a_k}$$

The value of C is given by:

$$C = \left.\frac{\cancel{(s+b_0)^2}\prod_i (s+a_i)}{\cancel{(s+b_0)^2}\prod_j (s+b_j)}\right|_{s=-a_0}$$

Since the sum of the residues must be equal to zero or one we find that

$$R_0 = \delta_{(p+1)(n+2)} - \sum_{k=1}^{n} R_k$$

where

$$\delta_{(p+1)(n+2)} = \begin{cases} 1 \text{ for } p+1 = n+2 \\ 0 \text{ for } p+1 \neq n+2 \end{cases}$$

C. One Third Order Pole:

From the equation:

$$\frac{(s+a)}{(s+b_1)^3(s+b_2)} = \frac{A}{(s+b_1)^3} + \frac{B}{(s+b_1)^2} + \frac{C}{(s+b_1)} + \frac{D}{(s+b_2)}$$

We find as before:

$$A = \frac{(a-b_1)}{(b_2-b_1)} \qquad D = \frac{(a-b_2)}{(b_2-b_1)^3}$$

then since $C+D=0$

$$C = -D$$

to find B both sides of the equation are multiplied by $(s+b_1)$ to obtain a new equation with only a second order pole. The new polynomial ratio is expanded into a partial fraction and the terms of this expansion are equated to similar ones from $(s+b_1)$ times the original expansion to give:

$$\frac{s+a}{(s+b_1)^2(s+b_2)} = \frac{A}{(s+b_1)^2} + \frac{B}{(s+b_1)} + \frac{F}{(s+b_2)}$$

Note: F could have been written

$$F = C + \frac{D(s+b_1)}{s+b_2}$$

but it is simpler to evaluate and use if it is taken from the new equation as indicated.

Both A and B are the same as in the above equation and F can be found easily:

$$F = \frac{a-b_2}{(b_1-b_2)^2}$$

then since $F+B=0$

$$B = -F$$

D. One High Order Pole:

The same technique can be extended to include all such functions with a single high order pole.

The equation to be solved is:

$$\frac{\prod_i (s+a_i)}{(s+b_0)^m \prod_j (s+b_j)} = \sum_{\ell=1}^{m} \frac{C_\ell}{(s+b_0)^\ell} + \sum_{k=1}^{n} \frac{R_{k1}}{(s+b_k)}$$

$i = 1, 2, \cdots, p$
$j = 1, 2, \cdots, n$
$p \leq (n+1)$

Defining

$$R_{k\ell} = \left. \frac{(s+b_k) \prod_i (s+a_i)}{(s+b_0)^{m-\ell+1} \prod_j (s+b_j)} \right|_{s=b_k}$$

The values of R_{k1} can be determined from this equation and the C's can be found from:

$$C_m = \left. \frac{\prod_i (s+a_i)}{\prod_j (s+b_j)} \right|_{s=b_0}$$

$$C_\ell = \delta_{(m+n-\ell)(p)} - \sum_{k=1}^{n} R_{k\ell}$$

$$\delta_{(m+n-\ell)(p)} = \begin{cases} 1 \text{ for } m+n-\ell = p \\ 0 \text{ for } m+n-\ell \neq p \end{cases}$$

for values of ℓ such that

$$1 < \ell < m-1$$

E. Two Second Order Poles:

More than one high order pole can be handled also:

$$\frac{1}{(s+b_1)^2(s+b_2)^2} = \frac{A}{(s+b_1)^2} + \frac{B}{(s+b_1)} + \frac{C}{(s+b_2)^2} + \frac{D}{(s+b_2)}$$

A and C are easily found. They are:

$$A = \frac{1}{(b_2-b_1)^2} \qquad C = \frac{1}{(b_1-b_2)^2}$$

Then multiplying both sides of the original equation by $(s+b_1)$ gives:

$$\frac{1}{(s+b_1)(s+b_2)^2} = \frac{A}{(s+b_1)} + B + \frac{C(s+b_1)}{(s+b_2)^2} + \frac{D(s+b_1)}{(s+b_2)}$$

The last two terms can be expanded by the same techniques to give four terms, the first two are:

$$C\left[\frac{(b_1-b_2)}{(s+b_2)^2} + \frac{1}{(s+b_2)}\right]$$

The coefficient of one for the $(s+b_2)$ term results because the numerator is one order less than the denominator. Therefore, the sum of the residues is one. There is only one first order pole term in the expansion so its coefficient is this residue of one.

The last term expands into

$$\frac{D(b_1-b_2)}{(s+b_2)} + D$$

This gives

$$\frac{1}{(s+b_1)(s+b_2)^2} = \frac{A}{(s+b_1)} + B + \frac{C(b_1-b_2)}{(s+b_2)^2} + \frac{C}{(s+b_2)} + \frac{D(b_1-b_2)}{(s+b_2)} + D$$

the coefficient of the second order term can be more easily evaluated directly. Calling this coefficient E gives

$$C(b_1-b_2) = E = \left. \frac{1}{(s+b_1)} \frac{(s+b_2)^2}{(s+b_2)^2} \right|_{s=b_2} \qquad E = \frac{1}{(b_1-b_2)}$$

The sum of the residues is zero so

$$A + C + D(b_1 - b_2) = 0$$

Therefore

$$D = -\frac{A+C}{(b_1-b_2)}$$

But

$$\frac{1}{(b_1-b_2)} = E$$

So

$$D = -(A+C)E = -2E^3$$

And the last coefficient is found to be

$$B = -D$$

Thus all the coefficients have been found by algebra, no differentiation was needed. The resulting equation can be reduced to

$$\frac{1}{(s+b_1)^2(s+b_2)^2} = E^2\left[\frac{1}{(s+b_1)^2} + \frac{E}{(s+b_1)} + \frac{1}{(s+b_2)^2} - \frac{E}{(s+b_2)}\right]$$

Where

$$E = \frac{1}{(b_1-b_2)}$$

These operations can be extended to higher order poles as required. With the basic rule about the sum of the residues, applied with usual algebraic manipulations, it appears that all differentiation can be avoided.

III Additional Properties of Residues

A. For a normalized polynomial with the same order numerator and denominator

$$P = \frac{s^n + a_1 s^{n-1} + \cdots + a_n}{s^n + b_1 s^{n-1} + \cdots + b_n}$$

the sum of the residues is $a_1 - b_1$

The proof can be obtained by observing that $\frac{P-1}{a_1-b_1}$ is a normalized polynomial with a numerator of one degree less than the denominator. The sum of the residues must be one. $P-1$ has the same residues as P and they must be (a_1-b_1).

B. If the numerator and denominator of a polynomial P are of the same order then the sum of the residues of $P + \frac{1}{P}$ will be zero. By (A)

$$\Sigma R(P) = (a_1 - b_1)$$

and

$$\Sigma R\left(\frac{1}{P}\right) = (b_1 - a_1)$$

then

$$\Sigma R(P) + \Sigma R\left(\frac{1}{P}\right) = 0$$

C. If the numerator of a polynomial is one degree higher than the denominator

$$P = \frac{s^{n+1} + a_1 s^n + a_2 s^{n-1} + \cdots + a_{n+1}}{s^n + b_1 s^{n-1} + b_2 s^{n-2} + \cdots + b_n}$$

then the

$$\Sigma R(P) = a_2 + b_1 - a_1 - b_2$$

This can be shown in a manner similar to A.

D. For a polynomial in which the numerator is $(n-1)$ larger than the denominator the first n coefficients of the numerator and the first n coefficients of the denominator are sufficient to determine the sum of the residues.

Conclusion

A theorem concerning the sum of the residues of the partial fraction expansion of a ratio of polynomials has been presented. It has been applied to the practial determination of the residues of several often-encountered expressions. In addition, other relations have been given which may prove useful in expansions of greater complexity.

Appendix

Theorem: For a normalized ratio of polynomials:

A. If the denominator is one degree higher than the numerator, the sum of the residues is one.

B. If the denominator is two or more degrees higher than the numerator, the sum of the residues is zero.

Proof: Let the polynomial ratio be written in the form:

$$P = \frac{s^{n-r} + a_1 s^{n-(r+1)} + \cdots + a_{(r+1)}}{s^n + b_1 s^{n-1} + \cdots + b_n}$$

Let s be written in the form:

$$s = Re^{j\theta}$$

All of the poles are near the origin so that as R approaches infinity the value of the polynomial is approaching

$$\lim_{(R\to\infty)} P \longrightarrow \frac{e^{-jr\theta}}{R^r}$$

and the closed loop integral

$$\lim_{(R\to\infty)} P \longrightarrow \int_0^{2\pi} PRe^{j\theta} d\theta = \begin{cases} 2\pi j & \text{for } r=1 \\ 0 & \text{for } r\leq 2 \end{cases}$$

therefore according to the residue theorem the sum of the residues will be

$$\Sigma R_k = \begin{cases} 1 & \text{for } r=1 \\ 0 & \text{for } r\leq 2 \end{cases}$$

COHERENT OPTICAL DATA PROCESSING

L. J. Cutrona, E. N. Leith and L. J. Porcello
The University of Michigan
Ann Arbor, Michigan

Summary

Coherent optical systems, which utilize the wave nature of light and the consequent diffraction phenomena, may often be used to supplement or even replace complex electronic equipment. Such systems are particularly adapted to the performance of certain linear mathematical operations, particularly those of an integral transform nature such as spectral analysis, convolution, auto- and cross-correlation, and matched filtering. The two-dimensional nature of optical systems, contrasted with the inherent one-dimensional nature of an electronic channel, allows a great reduction in equipment complexity for certain classes of operations.

This paper discusses the theory behind optical channels and filters as outlined above, and also illustrates simple multi-channel optical systems which can carry out representative operations.

Introductory Remarks

The use of optical techniques for processing of information has been frequently proposed. Both coherent and non-coherent optical systems have been proposed and built. These optical systems possess tremendous potential as data processors as a consequence of (1) the large storage capacity of photographic materials (2) the two-dimensional nature of the storage medium and (3) the ease with which wide bandwidth signals can be recorded and made static on photographic materials.

Despite these potentialities, optical systems have had limited fulfillment because of a number of severe problems, some of which have been overcome by the use of techniques to be described below. The advances in optical computing techniques which form the major content of this paper are the result of (1) the use of coherent optical techniques, and (2) the modification of the optical computer to have a multi-channel capability.

The properties of coherent optical systems are not original with the authors, although it is believed that some of these properties are used in novel fashion. The multi-channel use of an optical system also is not new. However, it is the belief of the authors that multi-channel operation prior to the work to be described herein has been considerably less versatile. The bulk of this paper, therefore will be devoted to the description of the configuration and to the properties of a multi-channel coherent optical signal processing system. Much of the versatility of the system will be seen to be a consequence of the fact that a coherent optical system is essentially a spectrum analyzer. A derivation of this result is given in Appendix I.

An important capability of one optical configuration to be described in this paper is the multi-channel evaluation of integrals of the form

$$I(y,t) = \int_{a(y,t)}^{b(y,t)} f(x,y,t)\, g(x,y)\, dx \qquad (1)$$

While important computations of this form are well known, it is perhaps useful to mention a few examples. This is done in Table I in which the specialized interpretations of the quantities in equation (1) are listed.

The importance of the optical technique which evaluates the integrals of the form given as equation (1) arises in part from the importance of problems such as those listed in Table I. However, of equal importance is the ease with which such optical computations can be made. Some remarks to this effect are the following:

(1) The recording of the functions $f(x,y,t)$ and $g(x,y)$ can be done readily. The photographing of intensity modulated cathode ray presentations or their equivalent is a direct means to convert to static form signal waveforms of very large bandwidths. While care is required in the design of the camera and film transport mechanisms when multi-channel operation is required, the process is essentially straight forward. Moreover, the staticising of the signals means that signals with bandwidths of the order of 100 Mc/sec can be recorded.

(2) A large number of independent channels can be evaluated simultaneously. The astigmatic

TABLE I SOME SPECIAL CASES OF THE FORM OF EQUATION (1)

E(y, t)	f(x, y, t)	g(x, y)	a(y, t)	b(y, t)
Fourier Transform	f(x)	$\sin 2\pi xy$ $\cos 2\pi xy$	$-\infty$*	$+\infty$*
LaPlace Transform	f(x)	$e^{-2\pi xy}$	0	∞*
Auto-Correlation	f(x)	f(x-y)	$-\infty$*	∞*
Cross-Correlation	f(x)	g(x-y)	$-\infty$*	∞*
Antenna Pattern	f(x) = antenna aperture distribution	$\cos\left(\frac{2\pi}{\lambda} xy\right)$, $y = \sin\theta$	Over antenna aperture	

*Obviously integration over an infinite interval cannot be performed. However, in many practical cases either f(x, y, t) or g(x, y) will vanish outside some interval. This replaces the infinite integral with appropriate finite limits.

optics described later keeps each channel separate so that at least 20 channels per mm of film can be accomodated.

(3) The coherent optical system behaves as a spectrum sorter.

A band pass filter becomes an appropriately placed aperture, a stop band filter becomes an obstacle in the optics. This permits simple sorting of spectra and permits easy processing in the frequency domain.

The exploitation of these characteristics is described in some detail in the sections which follow. In the sections which follow a single channel optical configuration which performs a two-dimensional spectrum analysis is described.

Following this discussion the modification of this configuration to produce a multi-channel one-dimensional spectrum analysis will be described. In a later section further characteristics of the multi-channel spectrum sorting capability of a coherent optical system will be developed, with emphasis placed on the use of optics to evaluate equation (2).

Coherent Optical Systems

In this section several coherent optical configurations and their characteristics will be described. The first configuration considered represents the state of the art before the work of the authors. In this configuration spherical optics only are used. This configuration gives a single channel, two dimensional spectrum sorting capability.

In each of the configurations considered later in this paper, astigmatic optics are used to achieve multi-channel operation, with one-dimensional spectrum sorting properties.

Single Channel, Two Dimensional Spectrum Analyzer

A coherent optical system without astigmatic elements such as that shown in Figure 1 produces a two-dimensional spectrum analysis in a configuration producing Franhaufer diffraction. The behavior of such a system can be described by

$$E(\alpha, \beta) = \underbrace{\iint}_{\text{over aperture defining A(x, y)}} A(x, y)\, e^{j(\alpha x + \beta y)}\, dxdy \qquad (2)$$

In this expression, α and β are direction cosines of the ray direction which produces a signal amplitude $E(\alpha, \beta)$ in the focal plane P_2 of lens L_2. A(x, y) specifies the transmission characteristics of a transparency in plane P_1 interposed between collimating lens L_1 and L_2. It may be real or complex.

Frequency filtering can be carried out very simply in plane P_2. Any obstacle at a point in P_2 defined by (α, β) is a stop band, any transparent region is a passband. Moreover, the use in P_2

of neutral density attenuating transparencies, with or without phase modifying characteristics, can be used to change either the magnitude or the phase of the spectral components in this plane.

It is evident from equation (2) that

$$\alpha = \beta = 0 \qquad (3)$$

corresponds to the zero frequency or average value of A(x, y) averaged over the aperture in plane P_1. It is further evident from (2) that the two space frequencies increase as α and β increase. The higher space frequencies (cycles per unit length) present in A(x, y) pass through plane P_2 at greater values of α and β. An aperture in plane P_2, therefore sets the upper limits on the details which can be transmitted beyond plane P_2.

If P_3 is the image plane for P_1, the image in this plane can be modified by processing in plane P_2. This modification is the result of the spectral components α, β which are allowed to impinge on plane P_3. Filtering in plane P_2 to select the frequency components allowed to pass to P_3 can be carried out to provide a number of functions. Prior work by O'Neill (Refs. 4, 6) has shown how a central obstacle at

$$\alpha = \beta = 0$$

removes the average level and increases the contrast of a photograph. He has also shown that the use of appropriate passbands in P_2 can be used to selectively emphasize signals in P_1. Specifically shape selection and noise reduction were demonstrated.

The Use of Astigmatic Optics To Achieve A Multi-channel Spectrum Analyzer

The work performed by the authors has concerned itself with a modification of the configuration of Figure 1 to that shown in Figure 2. This configuration differs from that of Figure 1 in that two cylindrical lenses L_4 and L_5 have been added.

The effect of the cylindrical lens L_4 is to produce focusing of rays leaving plane P_1 from a line, y = constant, to a corresponding line, y' = constant, in plane P_2. Similarly light passing through a line, y' = constant, in P_2 arrives at a corresponding line, y'' = constant, in plane P_3. Along such lines however, there will be a spectral decomposition of the structure. Specifically along a line

$$y' = \text{constant}$$

in plane P_2 will be found a spectral analysis of the information along the corresponding line

$$y = \text{constant}$$

in plane P_1. Similarly along a line

$$y'' = \text{constant}$$

in plane P_3 will be found a spectral decomposition of the light along a line

$$y' = \text{constant}$$

in plane P_2. The configuration of Figure 2, therefore is a multi-channel spectrum analyzer.

The multi-channel capability is easily demonstrated experimentally by placing an opaque obstacle with sides parallel to the x-axis in the P_1 plane. This will remove light from the corresponding intervals in P_2 and P_3. Moving this obstacle in the P_1 plane keeping its edges parallel to the x-axis causes a corresponding movement in the P_2 and P_3 planes. The edges of the corresponding areas in planes P_2 and P_3 are sharp and correspond to the resolution of the optical system.

The implication of this result is that the resolution capability of the optical system determines the number of channels per unit length along the y direction. Successful operation with about 12 channels per mm has been accomplished with expectations that this packing can be increased to at least 20 channels per mm. Using 35 mm film, therefore, with 24 mm of the film width active, approximately 250 to 500 independent channels of information are available. The number of independent channels can be increased over this value in proportion to the film width used.

The distribution of light amplitude along any line

$$y' = \text{constant}$$

in plane P_2 can be described by

$$E(y', \theta) = \int A(x, y)\, e^{j(2\pi/\lambda)\, x \sin\theta}\, dx \qquad (4)$$

where the integration with respect to x is carried out along the line

$$y = \text{constant}$$

in plane P_1, which corresponds to the line

$$y' = \text{constant}$$

in plane P_2.

In equation 4, λ is the wavelength of the light source in Figure 2, while θ is the angle in a plane normal to P_2, for

$$y' = \text{constant},$$

between the optical axis of the system and the deviation of the light. Since P_2 is the focal plane of lens L_2 the correspondence between the point x' and the angle θ having the amplitude given by equation 4 is

$$x' = F \tan \theta$$

where F is the focal length of lens L_2.

The remarks which have been made concerning the relation of the light distribution in plane P_2 to that in plane P_1 are pertinent also to the relation of the light amplitude distribution in plane P_3 relative to those in plane P_2 - namely: each line in P_3,

$$y'' = \text{constant}$$

has a light distribution along it whose amplitude is the spectral analysis of the light along the corresponding line

$$y' = \text{constant}$$

in plane P_2.

As contrasted to the remarks made with respect to Figure 1, the spectral decompositions made by the components shown in Figure 2 are one dimensional. As before, however, obstacles in plane P_2 (or P_3) become stop bands, while transparent regions become pass bands. Thus, modification of the spectral content of each line of A(x, y) is possible in plane P_2.

The possibility of pass bands and stop bands has already been mentioned. However, it is possible to modify the spectral composition both in amplitude and in phase by locating neutral absorption material and/or phase plates in plane P_2. Again since each channel is independent, the operations performed in each channel can be made independent if desired.

The versatility of the configuration of Figure 2 (or slight modifications thereof) will be illustrated by several examples - namely:

(1) the use of the configuration as a multichannel filter,

(2) the use of a modified configuration to evaluate integrals of the form

$$I(y, t) = \int_{a(y,\,t)}^{b(y,\,t)} f(x, y, t)\, g(x, y)\, dx \qquad (5)$$

Spectrum Analysis Experiments. Several experiments illustrating the spectrum analysis capability of coherent optical systems will be described in this section.

Let the configuration of Figure 2 be considered for the case that A(x, y) consists of a transparency as shown in Figure 3.

This transparency consists of a number of strips of alternately opaque and transparent lines. The various strips have varying space frequencies (cycles per unit length). The transmissivity across each strip has the form given by Figure 3b. It can be considered to be the superposition of an average or d-c, level plus a square wave.

In the P_2 plane of Figure 2, a spectral analysis of A(x, y) by lines will be found. The result of photographing the light amplitude distribution in plane P_2 with A(x, y) as shown in Figure 3 is given in Figure 4.

It will be noted that all channels have a recorded signal at the center. This corresponds to the average level of illumination emerging from plane P_1. The images in each strip show several lines on each side of the central image corresponding to the fundamental and to the harmonics of the signal in that strip. It will be noted that strips having lower space frequencies have spectral lines of closer spacing than do the strips with higher space frequencies.

Examination of plane P_3 in Figure 2 reveals an image of plane P_1. This is a consequence of performing a second spectral analysis on the light distribution in plane P_2. A photograph of this image (for only one channel) is given as Figure 5a. If an obstacle is placed in plane P_2 to remove the central image, the contrast between the black and light regions disappears, as in Figure 5b. This interesting behavior can be understood by recalling that the eye and any photographic recorder respond to light intensity, the square of the light amplitude. Before removal of the d-c component, the total light amplitude (bias plus square wave) was necessarily positive. Removal of the bias level however, causes the

amplitude of the light reaching plane P_3 to have both positive and negative amplitudes. Since the eye and/or photographic recorders can only sense the square of this amplitude, both positive and negative amplitudes appear alike. This is the reason that the contrast in P_3 disappears when the d-c level is removed by filtering. Imperfect symmetry of the positive and negative half-cycles and loss of high-frequency components which accentuate the corners tend to cause the residual effects, in particular the dark lines between the half cycles and the remaining slight contrast between positive and negative half-cycles.

A second interesting experiment results from allowing the central image and the first set of side-bands to pass beyond plane P_2. In this case the image in plane P_3 resembles the pattern in plane P_1 except that the amplitude variation is now sinusoidal in waveform rather than square. This is a consequence of the filtering which has selected the average level and the signal fundamental. The result of performing this experiment is shown in Figure 6b. If the central image and the second pair of sidebands on each side of the central image are selected, the image in P_3 consists of twice as many lines per unit length as are present in plane P_1 (Fig. 6c), but the amplitudes are sinusoidal instead of square. This corresponds to selection of the average value and the second harmonic. Similar results can be obtained by selecting the central image and the 3rd or higher pair of side-bands, verifying that these components arise from the spectral composition of A(x, y) of Figure 3.

A Versatile Multi-channel Optical Computer

Let attention now be directed toward a configuration suitable for providing multi-channel outputs of the form of equation 5. The configuration of Figure 2 is modified to that of Figure 7.

In this configuration lens L_6 and cylindrical lens L_7 are added and P_4 is placed in the focal plane of L_6. A transparency $[B_f + f(x, y)]$ is placed in plane P_1, while a transparency $[B_g + g(x, y)]$ is placed in plane P_3, where B_f and B_g are bias levels for these transparencies.

In the absence of filtering in plane P_2, an image of plane P_1 would be formed in plane P_3. This would illuminate the transparency in plane P_3 so that the light amplitude emerging to the right of plane P_3 will be the product

$$[B_f + f(x, y)]\,[B_g + g(x, y)] = B_f B_g + B_g f(x, y) + B_{gf}(x, y) + f(x, y)\, g(x, y). \tag{6}$$

Since the light amplitude in plane P_4 is a spectral analysis of the signal across the corresponding line

$$y'' = \text{constant}$$

in plane P_3, the light amplitude distribution in plane P_4 consists of a line-by-line Fourier transform of expression 6.

$$E(y''', x''') = \int [B_f + f(x, y''')]\,[B_g + g(x, y''')]\, e^{j(2\pi/\lambda)\, x \sin\theta} dx$$
$$x''' = F_6 \tan\theta \tag{7}$$

To convert equation (7) to the form of equation (5) it is necessary to set $x''' = 0$ ($\theta = 0$) in equation (5) i.e., to accept the signals along $x''' = 0$ in Figure 7. Unfortunately the term $B_g B_f$ being the product of two constants is itself a constant and puts energy into plane P_4 at $x''' = 0$. This seriously reduces the contrast in the output. Some advantages of filtering in the P_2 plane may now be made apparent. If an obstacle is placed in the P_2 plane at $x' = 0$, the d-c component, B_f, of the transparency in plane P_1 can be removed. This makes the light amplitude incident on plane P_3 proportional to f(x, y) without the bias level. Thus the first and third terms of the right hand side of expression 6 vanish.

A question now arises concerning the possible error in evaluating equation (5) due to the second term of the right hand side of expression 6. This term $B_g f(x, y)$ can only be troublesome if f(x, y) has a d-c component. In this case a component due to this d-c level will appear at $x''' = 0$ in plane P_4. The error, due to this term however, can be removed from the system output by recording both f(x, y) and g(x, y) on a common carrier frequency. If the functions resulting from using f(x, y) and g(x, y) to modulate a common carrier frequency are designated by f'(x, y) and g'(x, y), then $B_g f'(x, y)$ will produce no output at $x''' = 0$ in P_4 and it can be shown that

$$\left[\int f'(x, y)\, g'(x, y)\, e^{j(2\pi/\lambda)x \sin\theta} dx\right]_{\theta=0} = \left[\int f(x, y)\, g(x, y)\, e^{j(2\pi/\lambda)x \sin\theta} dx\right]_{\theta=0} \tag{8}$$

A proof of this equality is given in Appendix II.

It is thus established that the configuration of Figure 7 can be used for the multi-channel evaluation of integrals of the form of equation 5. The above discussion has omitted many of the additional flexibilities which may be incorporated into coherent optical processors. Through necessity, discussions of many of the attendant problems have also been omitted.

Concluding Remarks

Systems of the type discussed above offer the possibility of synthesis of transfer functions having independent phase and amplitude characteristics over a two dimensional region. It appears possible to synthesize otherwise complex comb filters in rather convenient form — that of a diffraction grating. The ease with which the second degree of freedom can be traded off for a multi-channel capability is appealing for certain applications.

Among the list of disadvantages, one must include the noise-like effects of film grain, perturbations in emulsion thickness, and the effect of spurious scattering of light at various points. Investigations which will dictate the ultimate limitations and practicability of optical computing or data handling channels are far from complete. However, the flexibility and inherent simplicity of optical channels appear to assure this technique a promising role in forthcoming amenable computing and data handling problems.

References

1. Gabor, D., "Microscopy by Reconstructed Wave-Fronts," Proc. Roy Soc. (London) 1949, p. 454.

2. Elias, P., Grey, D., and Robinson, D., "Fourier Treatment of Optical Processes." Optical Society of America, 42, No. 2 (1952) p. 127.

3. Rhodes, J., "Analysis and Synthesis of Optical Images," American Journal of Physics, 21 (1953) p. 337.

4. O'Neill, E., "Spatial Filtering in Optics," IRE Transactions of the Professional Group on Information Theory IT-2; No. 2, June 1956.

5. Cheatham, T. P., Jr., and Kohlenberg, A., "Analysis and Synthesis of Optical Systems," Part I of Tech. Note 84, BUPRL, March 1952.

6. O'Neill, E., "The Analysis and Synthesis of Linear Coherent and Incoherent Optical System," Tech. Note 122, BUPRL, Sept. 1955.

Appendix I

A coherent optical system has the inherent property that it will perform a spectral analysis of an inserted signal. This property is a result of the wave nature of the illumination. A one-dimensional argument is given below to serve as a physical derivation and demonstration of the spectrum analyzer feature.

If one illuminates a diffraction grating using a coherent light source, one observes a diffraction pattern which consists of a central image plus symmetrically positioned pairs of images which are off the optical axis. The angle θ between the axis and the nth order spectral line is given by the well-known relation

$$n\lambda = d \sin \theta$$

where λ is the wavelength of the incident illumination and d is the grating constant.

For the first order spectrum, $\lambda = d \sin \theta$. If one now considers a grating having a smaller value for d, the 1st order image appears at a larger value of the angle θ. Thus one can establish, for a given order spectrum, a correspondence between grating constant d and angle θ.

Let us now consider a more general case, which we shall discuss in terms of the geometry of Figure A-1.

For unit incident intensity at the plane of the transparency, the emergent light amplitude may be described by

$$E = \mathrm{Re}\left[\sqrt{A(x)}\, e^{j(\omega t + \phi)}\right]$$

where
- ω = radian frequency of the light source
- $A(x)$ = transmission of the transparency, a function of the independent variable x
- ϕ = some arbitrary phase angle.

Consider the function

$$f(x) = \sqrt{A(x)}\,.$$

For any angle θ, the amplitude in the P-plane is

$$E(t, \theta) = \mathrm{Re}\left[\int_0^{X_{max}} f(x)\, e^{j[\omega(t-(x \sin \theta/c) + \phi]}\, dx\right]$$

$$= \mathrm{Re}[I].$$

The time average of the intensity at this angle θ is given by

$$P(\theta) = \frac{1}{T}\int_0^T E^2\, dt = \frac{1}{T}\int_0^T \{\mathrm{Re}[I]\}^2\, dt.$$

But

$$\mathrm{Re}[I] = \int_0^{X_{max}} f(x) \cos\left(\omega t - \frac{\omega x \sin \theta}{\lambda} + \phi\right) dx$$

$$= \int_0^{X_{max}} f(x) \cos \frac{\omega x \sin \theta}{c}\, dx \cdot \cos(\omega t + \phi)$$

$$+ \int_0^{X_m} f(x) \sin \frac{\omega x \sin \theta}{c}\, dx \cdot \sin(\omega t + \phi).$$

If f(x) is continuous in every finite interval and if $\int_{-\infty}^{+\infty} |f(x)|\, dx$ converges, then at every point x, $(-\infty < x < \infty)$ where f(x) has a right-and-left-band derivative, f(x) can be represented by its Fourier integral. The cosine and sine integrals for f(x) are given respectively by

$$f(x) = \frac{2}{\pi}\int_0^\infty \cos \alpha x \int_0^\infty f(x') \cos \alpha x'\, dx'\, dx$$

and

$$f(x) = \frac{2}{\pi}\int_0^\infty \sin \alpha x \int_0^\infty f(x') \sin \alpha x'\, dx'\, d\alpha.$$

In the above expression, the terms

$$\frac{2}{\pi}\int_0^\infty f(x') \cos \alpha x'\, dx'$$

and

$$\frac{2}{\pi}\int_0^\infty f(x') \sin \alpha x'\, dx'$$

are the Fourier cosine and sine coefficients corresponding to any value of α, $0 \leq \omega < \infty$.

If the function f(x') is defined within the interval $0 < x' < x_m$, and is zero outside this interval, the cosine coefficient becomes

$$F'_c(\alpha) = \frac{2}{\pi}\int_0^{X_m} f(x') \cos \alpha x'\, dx'.$$

Letting $\alpha = \frac{\omega \sin\theta}{c}$, one has

$$F'_c(\alpha) = \frac{2}{\pi}\int_0^{X_m} f(x') \cos\frac{\omega x' \sin\theta}{c}\,dx'$$

$$= F_c(\theta).$$

This is precisely the same form as was demonstrated above in the expression for E. Similarly

$$F_s'(\alpha) = F_s(\theta) \cdot$$

Therefore

$$\int_0^{X_m} f(x') \cos\frac{\omega x' \sin\theta}{c}\,dx' = \frac{\pi}{2} F_c(\theta)$$

$$\int_0^{X_m} f(x') \sin\frac{\omega x' \sin\theta}{c}\,dx' = \frac{\pi}{2} F_s(\theta)$$

$$\therefore E = \frac{\pi}{2}\left\{F_c \cos(\omega t + \phi) + F_s \sin(\omega t + \phi)\right\} \cdot$$

Note that F_c and F_s are not functions of t.

$$\therefore P(\theta) = (\pi/2)^2\left[F_c^2(\theta) + F_s^2(\theta)\right]$$

It is therefore clear that a spectral decomposition of f(x) is displayed in the diffraction plane where a correspondence exists between spatial frequency content of f(x) and off-axis position for the diffracted images.

Now consider independently the aperture amplitude distribution function f(x). Assume that a Fourier transform of this amplitude function has been made. Since the function is viewed as a linear superposition of its various frequency components, the diffraction pattern displays a number of lines with appropriate position and intensity values. Each line is a diffraction limited image of the initial point source of light, and its width is determined by the system aperture. Therefore the spatial frequency resolution capability of the spectrum analyzer is determined by its aperture, or the spatial extent of f(x). In the more general case, $P(\theta)$ may be a continuous spectrum, but the resolution limit is still a diffraction limit. The diffraction pattern will display the d-c component of f(x) as the central image and corresponding to each angle θ there will appear a source image whose intensity is proportional to the sum of the square of the Fourier coefficients for the frequency corresponding to that angle.

Appendix II

It is the purpose of this appendix to derive equation (8) of the text – namely that

$$\left\{\int_{-\infty}^{\infty} f(x)\, g(x)\, e^{-j(2\pi/\lambda)x \sin\theta}\,dx\right\}_{\theta = 0}$$

$$= \left\{\int_{-\infty}^{\infty} f'(x)\, g'(x)\, e^{-j(2\pi/\lambda)x \sin\theta}\,dx\right\}_{\theta = 0} \qquad \text{II-1}$$

where f'(x) and g'(x) are related to f(x) and g(x) by virtue of having had their spectra translated by ω_o radians per second.

Specifically if $F(\omega)$ and $G(\omega)$ are the Fourier transforms of f(x) and g(x) while $F'(\omega)$ and $G'(\omega)$ are the transforms of f'(x) and g'(x), then

$$F'(\omega)\begin{cases} = F(\omega-\omega_o) & \text{for } \omega > 0 \\ = F(\omega+\omega_o) & \text{for } \omega < 0 \end{cases} \qquad \text{II-2}$$

$$\begin{aligned} F'(\omega) &= 0 && \text{for } |\omega| < \omega_o \\ G'(\omega)&\begin{cases} = G(\omega-\omega_o) & \text{for } \omega > 0 \\ = G(\omega+\omega_o) & \text{for } \omega < 0 \end{cases} \\ G'(\omega) &= 0 && \text{for } |\omega| < \omega_o \cdot \end{aligned} \qquad \text{II-3}$$

If the value of θ in equation II-1 is set equal to zero then it will be sufficient to show that

$$\int_{-\infty}^{\infty} f(x)\, g(x)\,dx = \int_{-\infty}^{\infty} f'(x)\, g'(x)\,dx \cdot \qquad \text{II-4}$$

By Parseval's Theorem, one can write

$$\int_{-\infty}^{\infty} f(x)\, g(x)\,dx = \int_{-\infty}^{\infty} F(\omega)\, G(-\omega)\,d\omega \qquad \text{II-5}$$

and

$$\int_{-\infty}^{\infty} f'(x)\, g'(x)\,dx = \int_{-\infty}^{\infty} F'(\omega)\, G'(-\omega)\,d\omega \cdot \qquad \text{II-6}$$

It will be shown that the right hand side of equations II-5 and II-6 are equal. Then the left hand sides of these equations will also be equal and equation II-4 will be established, as will also equation II-1.

It is convenient to divide the interval of integration of the right hand side of II-6 into three

intervals — namely from $-\infty$ to $-\omega_o$, from $-\omega_o$ to ω_o, and from ω_o to $+\infty$. Since both $F'(\omega)$ and $G'(\omega)$ vanish for $|\omega| < \omega_o$ only two of the integrals need be evaluated. If the appropriate forms for $F'(\omega)$ and $G'(\omega)$ from II-2 and II-3 are used in II-6 the result is

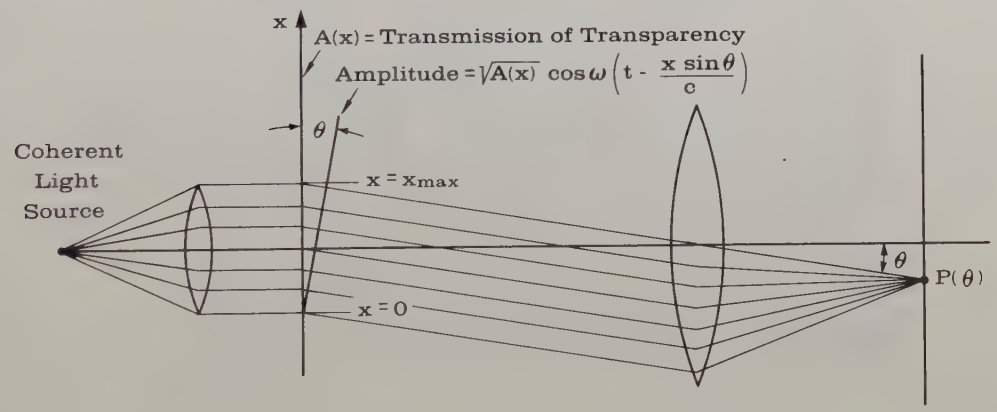
$$\int_{-\infty}^{\infty} F'(\omega)\, G'(-\omega)\, d\omega = \int_{-\infty}^{-\omega_o} F(\omega+\omega_o)\, G(-\omega-\omega_o)\, d\omega + \int_{\omega_o}^{\infty} F(\omega-\omega_o)\, G(-\omega+\omega_o)\, d\omega. \qquad \text{II-7}$$

By change of variable of integration the right hand integrals in II-7 can be written as

$$\int_{-\infty}^{\infty} F'(\omega)\, G'(-\omega)\, d\omega = \int_{-\infty}^{o} F(Z)\, G(-Z)\, dZ + \int_{o}^{\infty} F(Z)\, G(-Z)\, dZ = \int_{-\infty}^{\infty} F(Z)\, G(-Z)\, dZ.$$

This is the desired result.

The preceding work was supported by the U. S. Army and U. S. Air Force.



A-1 FRAUNHOFER DIFFRACTION GEOMETRY

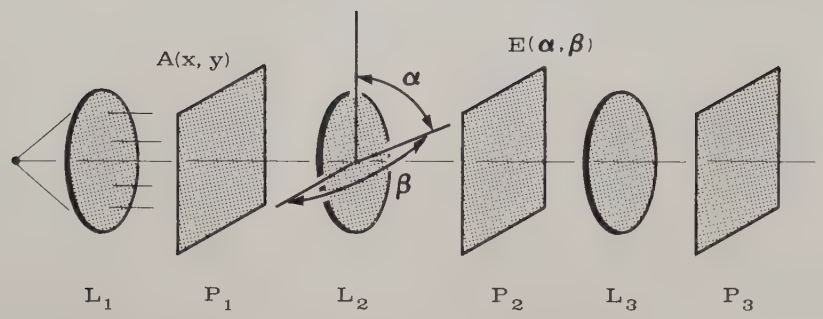


FIG. 1 CONFIGURATION PRODUCING FRAUNHOFER DIFFRACTION IN TWO DIMENSIONS

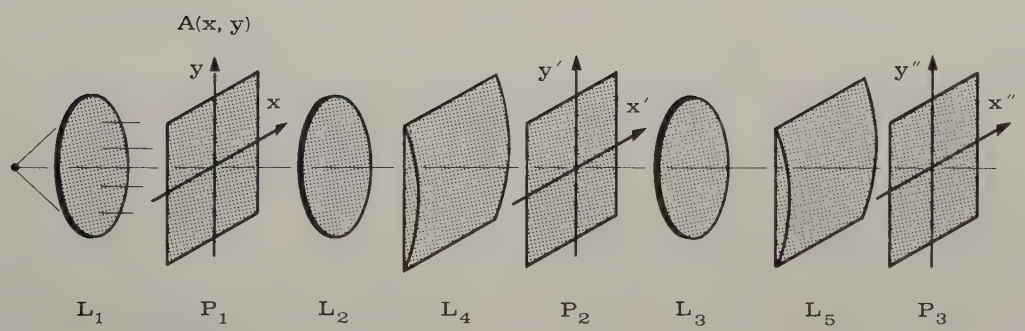
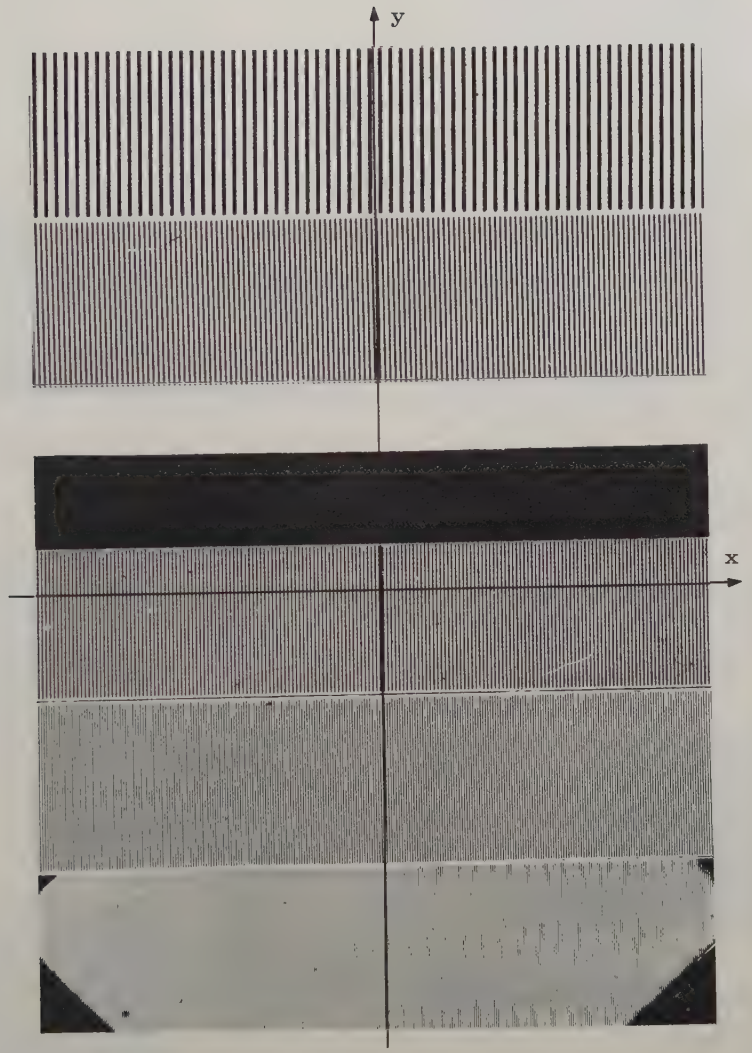


FIG. 2 MODIFICATION OF TWO-DIMENSIONAL CONFIGURATION TO INTRODUCE MULTI-CHANNEL CAPABILITY



a) Multi-channel diffraction grating.

b) Transmission function of grating for one channel.

FIG. 3

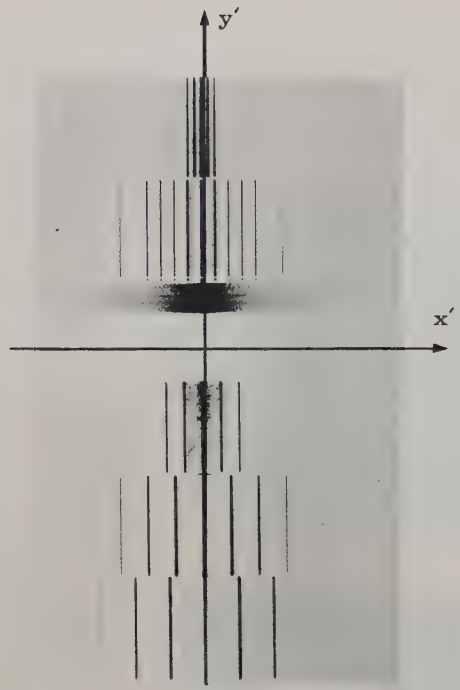
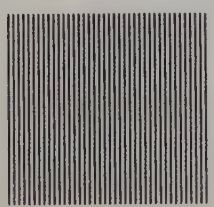


FIG. 4 DIFFRACTION PATTERN PRODUCED BY THE MULTI-CHANNEL DIFFRACTION GRATING

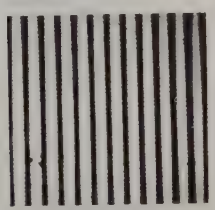


a) Square-wave diffraction grating.

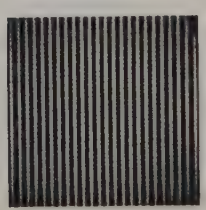
b) Image of square-wave grating after bias removal.

FIG. 5

a) Square-wave diffraction grating.



b) Image of square-wave grating when filter transmits d-c and fundamental only.



c) Image of square-wave grating when filter transmits d-c and 2nd harmonic only.

FIG. 6

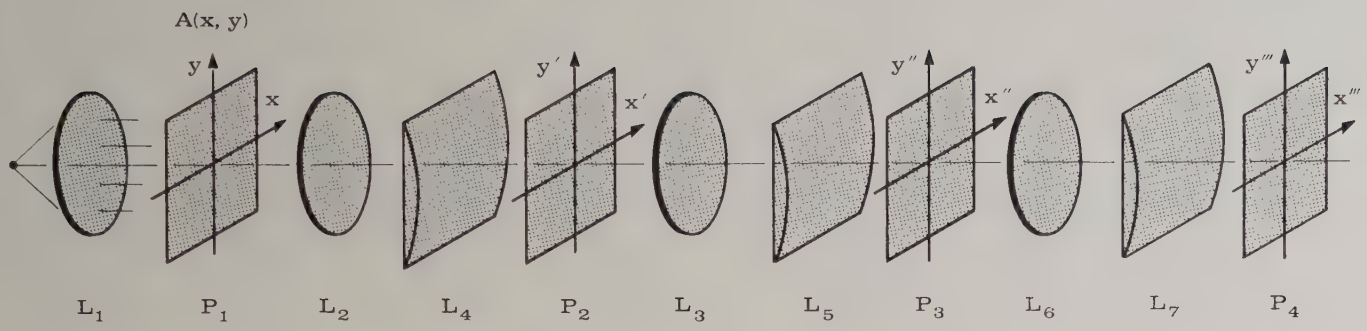


FIG. 7 ARRANGEMENT SHOWING GENERATION OF CONSECUTIVE TRANSFORM PLANES

POLE DETERMINATIONS WITH COMPLEX-ZERO INPUTS

John A. Brussolo
Beckman Instruments, Inc., Berkeley Division
and
University of California
Berkeley, California

Summary. A method for the determination of the characteristic pole locations in simple systems and components is discussed. The techniques introduced are extended to include more complex systems. The method consists of applying a signal to the system from a "complex-zero-generator" and then observing the system output response on an oscillograph. The system pole locations are determined when the system output goes through a null, which occurs when the zeros generated by the "complex-zero-generator" cancel the system poles. Tests on experimental systems indicate that the pole locations can be determined accurately and rapidly for a wide variety of systems.

For testing purposes, a "complex-zero-generator" was built using two square-wave generators with a timed delay between them. This generator gives a signal which can contain complex zeros anywhere in the s-plane. This signal is then applied to a number of systems containing several different relative pole orientations. The system outputs, as photographed on the oscillograph, are studied to develop rules and procedures to determine the pole locations. The rules are then applied to a number of "unknown" systems in order to determine their applicability and a study is made of the errors and limitations involved.

The results indicate that the method described is readily applicable to many systems and that, in many instances, the pole locations are determined more accurately and more rapidly than can be done through the use of a steady-state frequency analysis.

Introduction

A novel method of determining the characteristic pole locations in simple systems and components has been presented recently by George G. Lendaris and Otto J.M. Smith in a paper entitled "Complex-Zero Signal Generator for Rapid System Testing".[1] A brief exposition of this work is followed by an extension of the techniques described to more complex systems.

The method employed by Lendaris and Smith differs radically in its approach to the problem from the conventional procedures of making a steady state frequency analysis or a transient response analysis. Use is made of a "complex-zero-generator", (CZG). This signal is used as an input to the system, and the system output observed as an oscillographic trace. When the zeros generated by the CZG are placed so as to cancel the poles of the system being tested, a part of the system output will go through a null. At this point, the characteristic pole locations are given by the known zero locations determined by the CZG calibration.

Complex-Zero-Generator

It has been shown in the paper mentioned previously that complex zeros can be generated by using a double-step time function as shown in Fig. 1. This function has a pole at s = 0 and zeros where

$$a + b\epsilon^{-sT} = 0.$$

If we denote the s-plane locations of the roots of this input function by $s = \alpha + j\beta$ we find that the function has an infinite column of zeros in the s-plane where:

$$\alpha = \frac{1}{T}\ln\frac{b}{a} \quad ,$$

$$\beta = \pm \frac{(2n+1)\pi}{T} \qquad b>0$$

$$n = 0,1,2,\ldots.$$

$$\beta = \pm \frac{2n\pi}{T} \qquad b<0$$

The s-plane locations of these zeros are shown in Fig. 2 along with the corresponding time function. We see that by varying the b/a ratio in sign and magnitude we can generate real or complex zeros in any part of the s-plane.

A generator to provide double-steps was constructed and, in block diagram form, appears as shown in Fig. 3. The main elements consist of two square wave generators, one of which is controlled by the other. The slave SWG delivers a square wave delayed by a variable time T after the master SWG. The two signals are added and produce the output wave shown in Fig. 3. Since the system output is periodic, if the square wave period is made long in comparison to the time that it takes the system transient

to die out, the system output can be continuously displayed on an oscillograph screen.

Pole Determinations for Simple Systems

Lendaris and Smith describe the method of operation necessary to find the characteristic pole locations for simple systems consisting of one real pole, a complex pole-pair, or one real pole plus one complex pole-pair.

Consider first a simple system consisting of a single complex pole-pair. The s-plane pole-locations for this type of system along with the single-step response are shown in Fig. 4. If we use the CZG to generate a pair of complex zeros as an input to the system, the system response will consist of the normal single-step response up to the time of the introduction of the second step. After the second step is applied the system response is composed of the sum of the single-step response and the response due to the second step. If, when the second step is introduced, the delay time T and the amplitude ratio b/a have been adjusted to give zeros exactly at the system pole locations, after the time T, the response will be deadbeat (i.e. a constant) since the poles have been removed from the response by cancellation with zeros. This type of deadbeat response is shown in Fig. 5.

It can be shown that, for a system consisting of a single complex pole-pair, the second step must occur in time at exactly the time of the peak of the first overshoot in the single-step response. A practical procedure would then involve using a dual-trace oscillograph which presents both the system response and the double-step input. The second step delay time T is adjusted until the second step occurs at the time of the first overshoot peak. Then the amplitude ratio b/a is adjusted to give a deadbeat response. The readings of the CZG dials then give the locations of the complex zeros which are the same as the system poles.

For a system with a complex pole-pair plus a real pole, the procedure to be used consists of nulling out first the complex pole-pair and then the real pole. With the complex pole-pair nulled out, the response observed is only that due to a single real pole and with the real pole nulled, the response is that due to the complex pole-pair alone. These responses are shown in Fig. 6 along with the single-step response.

Pole Determinations for Complex Systems

Extension of the methods discussed to more complex systems follows from the consideration that when one pole or complex pole-pair is nulled out the response remaining has the characteristics of the remaining poles only. Thus, if the separate responses for the real poles or complex pole-pairs in the system are different enough so that the eye can distinguish them, they can be nulled out separately, one-by-one. However, if the system poles are spaced closely together, the eye cannot distinguish as well the separate responses, and a different procedure must be adopted.

The approach used here was to rely on direct experimental observation of as many different systems as was practicable. By looking at the responses obtained in a variety of cases and analyzing these responses in several different manners, some general rules and procedures for investigating complex systems have been formulated.

Several systems were built and responses photographed for inputs containing different zero locations. The system which proved to be the most valuable consisted of a representation on an electronic analog computer.

The systems used were restricted to those having two complex pole-pairs spaced relatively closely together. In addition, the systems tested were lightly damped, i.e. the ratio ω/σ for each pole pair was made high, approximately 10/1. The reason this was done was to obtain several cycles of oscillation in the response in order to help develop criteria for finding the poles. Fig. 7 shows the pole locations in the s-plane for the seven types of systems studied.

The procedure followed was to set up each system individually and then, using the CZG to provide inputs to the system, to vary the second step time delay and amplitude ratio to give complex zeros located in various spots in the region of the poles. The response to each of these inputs was photographed and then the responses for each system were grouped together and photographed again. These appear as Figures 8 through 14. The responses are shown, placed in the s-plane. Each picture represents the response of the system obtained by using the CZG to place a zero at the picture location in the s-plane. The system pole locations can be seen in each figure as black crosses underneath a response photograph. The

figures show only the upper half of the left-half s-plane for economy.

Direct examination of Figs. 8 - 14 with the aid of a transient vector diagram analysis[2] enables us to develop rules for finding the characteristic pole locations in systems of this type. Before stating the rules developed, the main results of the analysis of Figs. 8 - 14 can be summarized as follows:

1. We can tell that the system has at least two complex pole-pairs by observing the first undershoot in the single-step response. The amplitude of the first undershoot is always greater than that which would be obtained with a single complex pole-pair. This is a consequence of the phases of the residues at the poles being about 180° out of phase. In general, one of the transient vectors is rotating faster than the other, and will catch up with and reinforce the slower one.

Also, in the single-step response, if measurements are made of the times between successive pairs of overshoot peaks it will be found that the frequency of the response will not be a constant but will vary for at least the first part of the response curve.

2. When a zero is placed exactly on top of one pole, the system response is due entirely to the other pole. This could be expected since the zero cancels the pole on which it lies and leaves only the other pole in the system.

3. We can always tell when the zero is outside the pole range and in which direction. As can be seen in Figs. 8-14, the responses obtained for a zero placed somewhat above the higher-frequency pole and for a zero placed somewhat below the lower-frequency pole have characteristics that set them apart from the responses obtained for a zero in between the poles. The response for a zero placed above the higher-frequency pole has an apparent resonant frequency approximately that of the lower-frequency pole and vice-versa.

4. We can always tell when the zero is inside the pole range. Many changes occur in the response for zeros placed in this region. It can be seen, by looking at Figs. 8 - 14, that the most characteristic responses are obtained for a zero placed, not only in between the poles, but on the $j\omega$-axis. Therefore, this would be a logical place to start the testing procedure on an unknown system.

5. We can find a point approximately halfway between the poles in terms of ω. It can be shown that this particular point is a unique one and is primarily the factor which enables us to find the pole locations accurately. As the zero is moved down the $j\omega$-axis from above the higher pole, when the zero reaches the region of the mid-frequency point, the response shows the beginning of a "bump" between the first and second overshoots. This is evident in Figs. 8 - 14 for all seven systems.

6. We can sometimes find a point approximately halfway between the poles in terms of σ; however, this point is not as well-defined as the one on the $j\omega$-axis.

7. The lower-frequency pole is generally much easier to find than the higher-frequency pole, since the responses for zeros placed in the region of the lower-frequency pole change rapidly for small changes in zero location.

8. Larger changes in the system response are obtained for smaller changes in zero location when the zero is close to the $j\omega$-axis.

Rules for Finding the Pole Locations

Based on an analysis of Figs. 8 - 14, three separate methods have been developed to find the characteristic pole locations for systems of this type. An additional method can be employed by properly averaging the results of two of the basic methods. Only one method gives the exact coordinates of the poles; the others give an approximation to the pole locations. It was felt that the approximate methods should be developed since they are easier to use and still give good results.

The first step in any method is to ascertain that the system does consist of two complex pole-pairs. Quite often this information can be deduced from the single-step response by looking at the amplitude of the first undershoot or by the presence of beat notes in the oscillations. If neither of these criteria is present, the operator of the CZG would try to find a null assuming a single complex pole-pair. When this cannot be achieved, it would then be assumed that there are at least two complex pole-pairs. It should be noted that the absence of a single exponential envelope decay or the presence of beat notes can give valuable clues to the relative orientation of the poles and thus simplify the task of finding their locations.

Method I. (See Fig. 15.)

1. Use a b/a ratio of 1.0 (i.e. $\alpha = 0$) on the CZG and vary T from low values to high values. This places a zero on the jω-axis and brings it down the axis from high frequencies to low frequencies. For β (the zero coordinate) above the higher pole, the response looks like that due to a single complex pole-pair. As β is made smaller, eventually a point is reached where the response departs from that due to a single complex pole-pair. At this point the response shows a "bump" starting to form between the first and second overshoots. When this occurs the zero is approximately halfway between the poles in terms of ω. The zero is stopped here and its location recorded from the CZG dial readings. Call the coordinates of this point: $(0, \omega_m)$.

2. T from the CZG (i.e. β) is now kept constant and b/a is varied toward 0 from 1.0. Now the generated zero is moving to the left in the s-plane along a line of constant frequency approximately halfway between the poles. As the zero is moved to the left, the response will begin to straighten out again to that of a single complex pole-pair. When this occurs, and where the decay rate is the same as it was for the response with zero placed above the higher-frequency pole in step 1., the zero is approximately halfway between the poles in terms of σ. The zero is stopped here and its location recorded from the CZG dial readings. The coordinates of this point we will call (σ_m, ω_m).

3. We now have the σ and ω approximately halfway between the poles. We can proceed to find one of the poles exactly and since the lower-frequency pole is easier to find because of the rapid response changes for a zero in its vicinity, we choose to find this one. To accomplish this, we vary T on the CZG in the proper manner to move the zero down from its midpoint position in small steps. At each step the b/a ratio on the CZG is varied to sweep a range of the s-plane. When the response obtained is that due to a single complex pole-pair only, the zero is sitting exactly on top of the lower-frequency pole. Call the coordinates of this point (σ_1, ω_1).

4. We have now found one pole and we can calculate the location of the higher-frequency pole, whose coordinates we will call (σ_2, ω_2). The higher-frequency pole coordinates are given by:

$$\sigma_2 = 2\sigma_m - \sigma_1 \quad ; \quad \omega_2 = 2\omega_m - \omega_1$$

Method II.

The procedure used in this method is the same as that used in Method I. However, here the midpoint coordinates (σ_m, ω_m) are ignored since they are only approximate and both the lower-frequency pole and the higher-frequency pole are found separately.

Method III.

This is the most accurate of the three methods employed and correspondingly takes more time and care to use. In this method, the midpoint between the poles is found as before but it is used as a reference point only, to place the zero in the vicinity of the poles. From here the zero is moved down or up and across the s-plane as in step 3 of Method I, and the response observed on the oscillograph. As the zero location is moved and the response changes, the horizontal gain control on the oscillograph is varied in order to allow comparison by eye of the times between the overshoot peaks. This is shown in Fig. 16.

When the zero is anywhere in the s-plane except on top of one of the poles, the frequency of the output will be a combination of the two pole-frequencies. This means that the time between successive overshoot peaks will not be constant but will vary, either increasing or decreasing for successive peaks. However, when one pole is cancelled by a zero, the response will contain only one frequency and the time between successive overshoot peaks will be a constant. This is relatively easy to check on an oscillograph screen. Thus the procedure is to get the zero from the midway point to the vicinity of a pole where the response looks like that due to a single complex pole-pair. At this point the times between successive peaks are measured on the oscillograph screen and the zero is moved slightly up and down and right and left until the measured times are equal. At this point the zero is exactly on top of the pole and the pole coordinates are determined from the CZG dial readings. This process is repeated for the other pole.

Errors and Limitations

There are several types of errors involved in the use of any of the methods described. Perhaps the largest er-

ror involved in all but Method III is the fact that finding a pole position depends on the ability of the eye to recognize a pure damped sinusoid. This can be extremely difficult under certain conditions, such as having the pole-pairs located very closely together as shown in Fig. 8. This is why Method III gives more accurate results than the other methods. It uses a procedure that allows one to detect differences very readily. Another large error which may occur when using the first two methods is in the determination of the σ-coordinate. For certain systems the error in the σ-coordinate can be as large as 100% unless Method III is used.

To check the accuracy of the methods developed, pole determinations were made on thirty-five different systems with random pole-pair spacing. The procedure followed was to set up a system on an electronic analog computer where each complex pole-pair was generated in a separate section and then the sections were cascaded. All three of the methods were employed, in sequence, to determine the system pole locations. The exact pole locations were also determined for each section separately using the CZG, since the method is extremely accurate for single complex pole-pairs.

The pole locations determined through the use of each of the three methods were compared with the true pole locations and the errors calculated. This was done for a rather small sample of thirty-five systems but the results show a good correlation. It should be noted that the types of systems tested included eleven different systems which had an ω/σ ratio of about 1.2 which would indicate that the methods described are valid for a wide range of the ω/σ ratio. This is true even though the empirical work used to derive the rules for locating poles was done for ω/σ ratios of about 10 to 1.

An examination of the errors involved in the determination of the pole locations for these systems shows that for Methods I and II, errors of up to 10% in the ω-coordinate and 16% in the σ-coordinate were obtained.

However, errors were obtained as high as 50% in the pole-to-pole spacing in the ω-direction and up to 200% in the pole-to-pole spacing in the σ-direction. The superiority of Method III is demonstrated by maximum errors of 2% and 5% in the ω-coordinates and σ-coordinates respectively. In addition, the maximum errors obtained for Method III in the pole-to-pole spacings were 2% and 9% in the ω-direction and σ-direction respectively. Thus, it is apparent that if a person desired accuracies of 15%, for example, in the actual pole positions, he could use Method I; but if he desired the same percentage of accuracy in the measurement of the pole-pair spacing he would need to use Method III.

We should note, before leaving the subject of errors in these methods of pole locations, that in general, several cycles of oscillation should be available to observe in the response. The reason for this is that Method III depends on having at least three overshoot peaks and this method is the most accurate. This means that better results are obtainable with lightly-damped systems than with heavily-damped systems. The errors obtained however, include systems with relatively heavy damping and the maximum errors result from these heavily damped systems. Therefore, for lightly damped systems we would expect the range of errors to decrease and we could predict much better results.

Conclusions

The results of the analysis and experimentation carried out indicate that the methods outlined are sufficient to enable the determination of pole locations in systems more complex than have been originally described. The procedures developed have their limitations of course; although it will be found that they cover a large number of systems or networks that are generally encountered. Few systems have more than two dominant complex pole-pairs in addition to two or three real poles.

It is felt that the techniques presented here can be extended to include some of the less common complex systems. In addition, there is room for improvement in the procedure itself in regard to speed, simplicity, and accuracy of operation. A more rigorous approach through the development of the associated mathematics would be highly desirable.

For the types of systems treated, this method of cancelling system poles with generated zeros offers relief from the tedious process of making a steady-state frequency analysis with or without a transient response analysis to obtain the system equations. In this respect it should be noted that, using the methods described, the pole locations for systems having up to two complex pole-

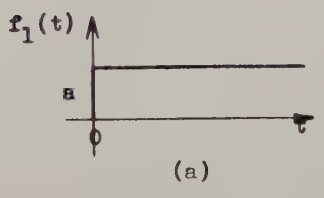
pairs can be found in from five to ten minutes by a moderately experienced person. This is less than one-third of the time required to make a steady-state frequency analysis for a comparable system.

The underlying simplicity of the pole-cancellation technique promises more general use of this method as further developments are made.

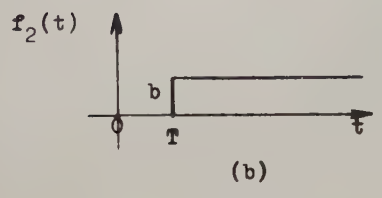
References

1. Lendaris, George G., and Smith, Otto J.M., "Complex-Zero Signal Generator for Rapid System Testing", Transactions Paper, No. 58-1027, American Institute of Electrical Engineers, New York, paper presented at the AIEE Pacific General Meeting, Sacramento, Calif., August 19-22, 1958.

2. Smith, Otto J.M., Feedback Control Systems, New York, Mc Graw Hill Book Co., 1958, p. 25.

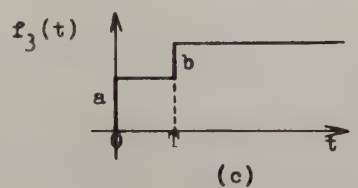


(a)



(b)

$$F_1(s) = \mathcal{L}\left[f_1(t)\right] = \frac{a}{s} \qquad F_2(s) = \mathcal{L}\left[f_2(t)\right] = \frac{b}{s}\epsilon^{-sT}$$



(c)

$$F_3(s) = \mathcal{L}\left[f_3(t)\right] = \mathcal{L}\left[f_1(t)+f_2(t)\right] = \frac{a}{s} + \frac{b}{s}\epsilon^{-sT} = \frac{1}{s}(a+b\,\epsilon^{-sT})$$

Poles: $s = 0$ Zeros: $a + b\,\epsilon^{-sT} = 0$

Fig. 1. Generation of Complex-Zeros

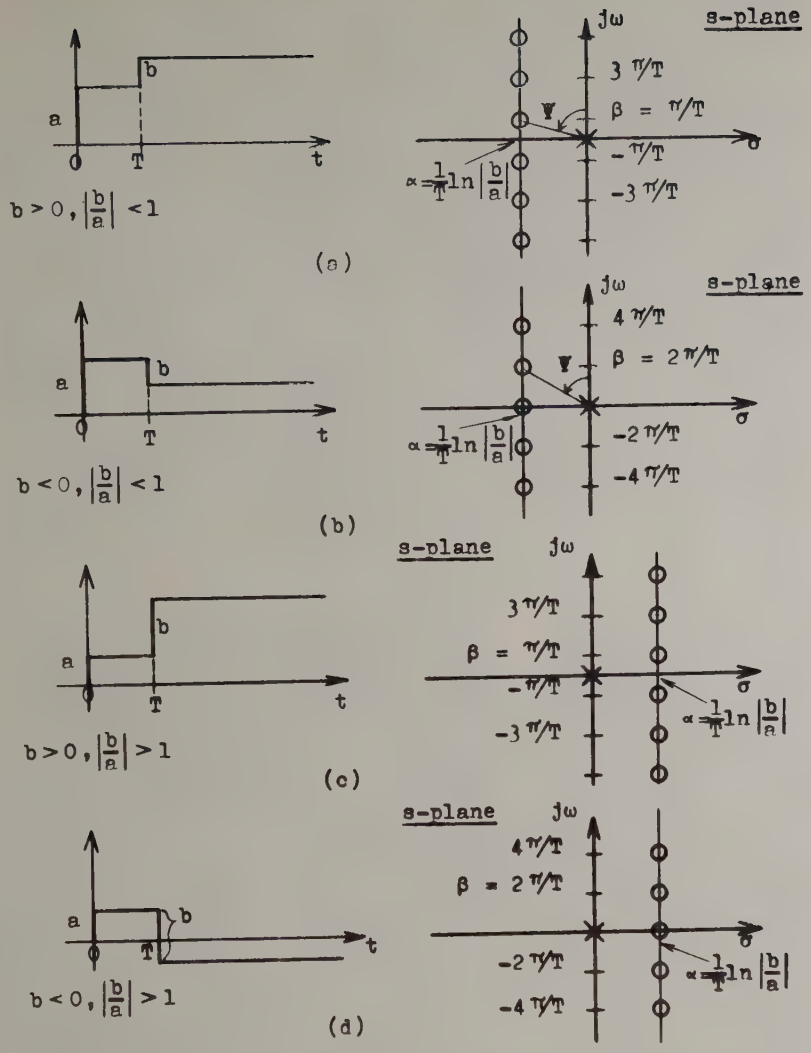


Fig. 2. S-plane Pole-Zero Locations For Double Steps In The Time Domain

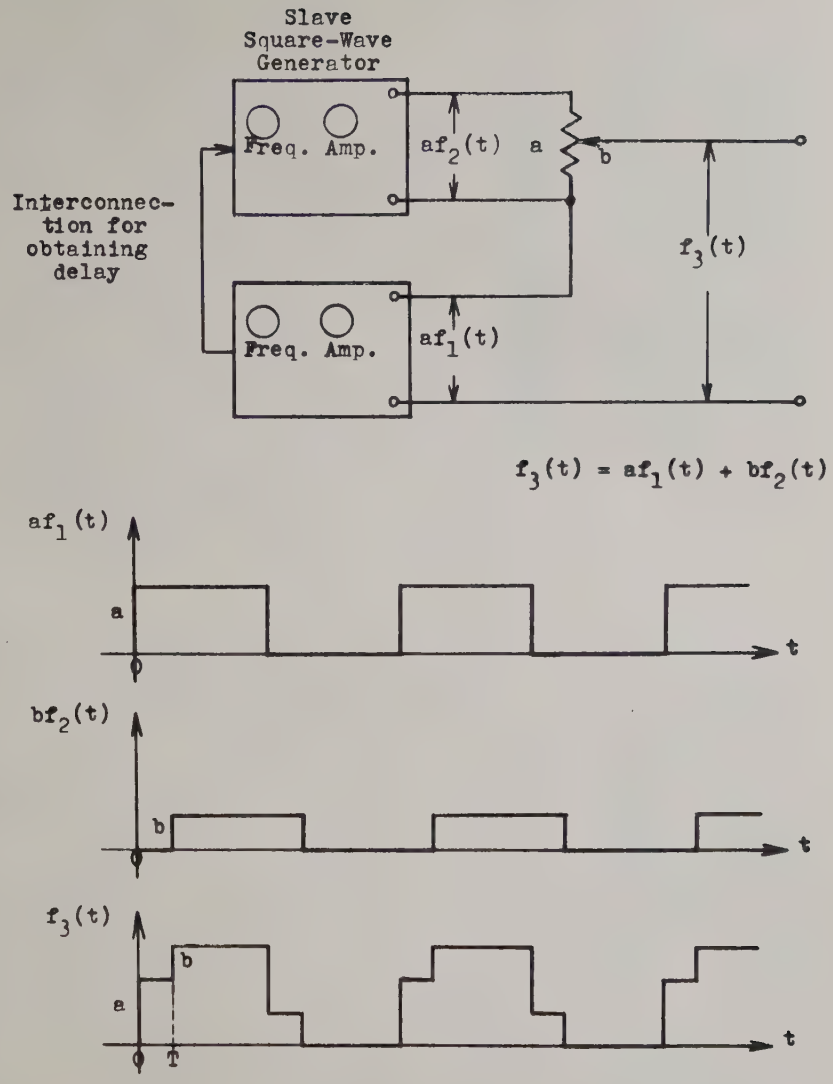


Fig. 3. Complex-Zero Generator

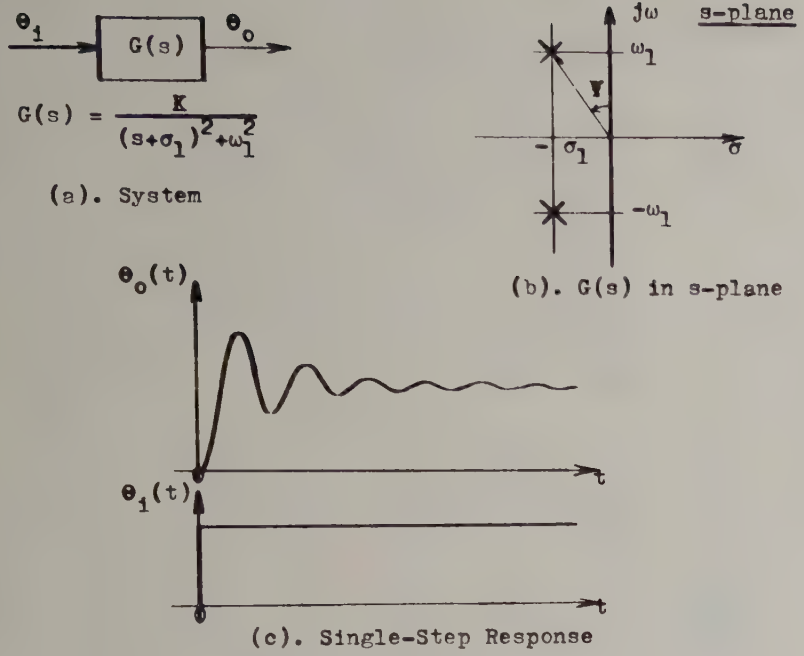


Fig. 4. Single Complex Pole-Pair System

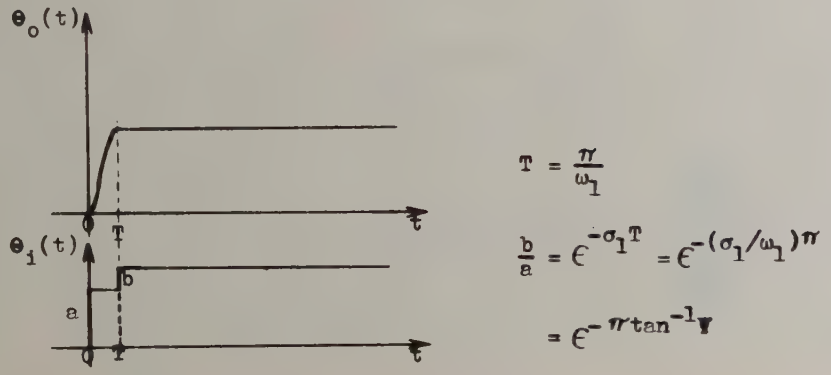
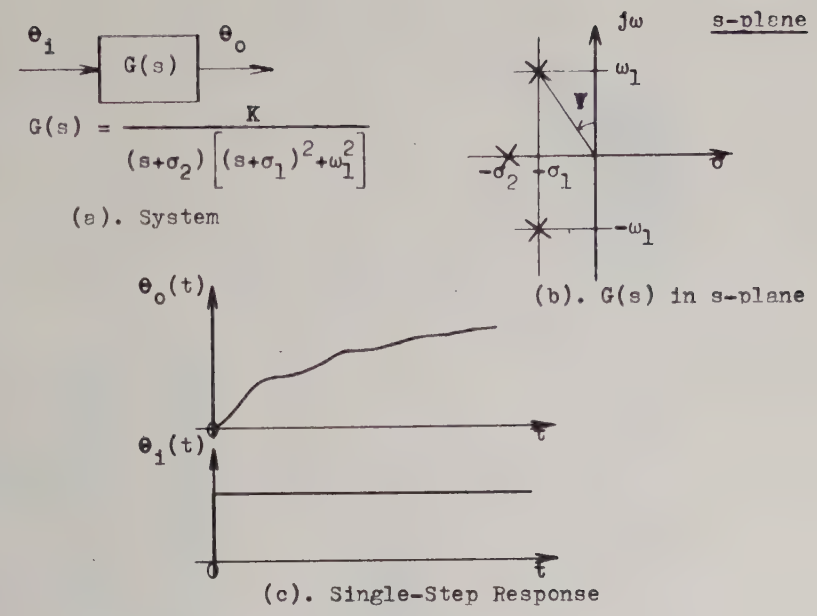


Fig. 5. Deadbeat Response



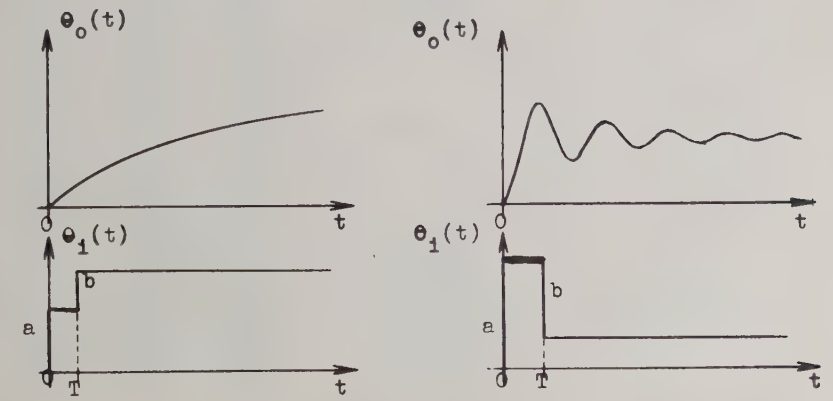


Fig. 6. Single Complex Pole-Pair Plus One Real Pole

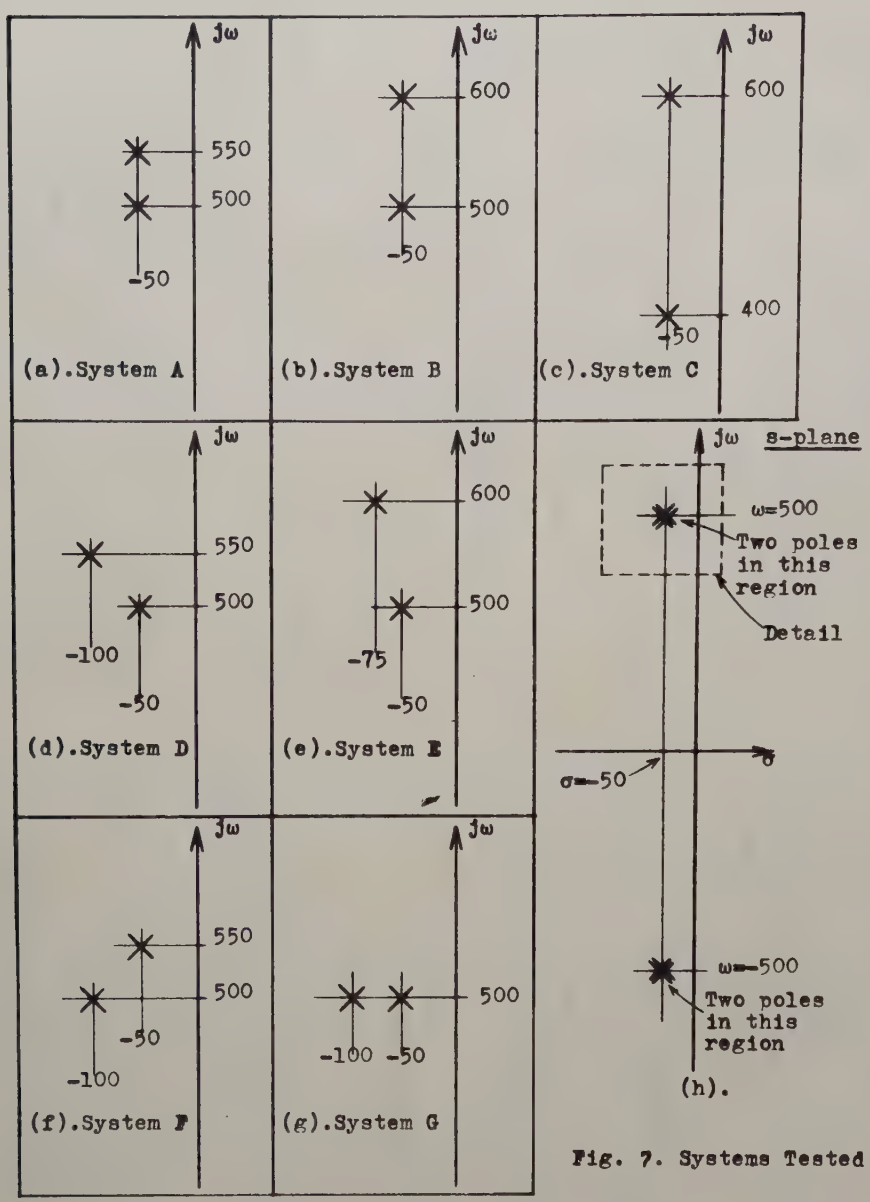


Fig. 7. Systems Tested

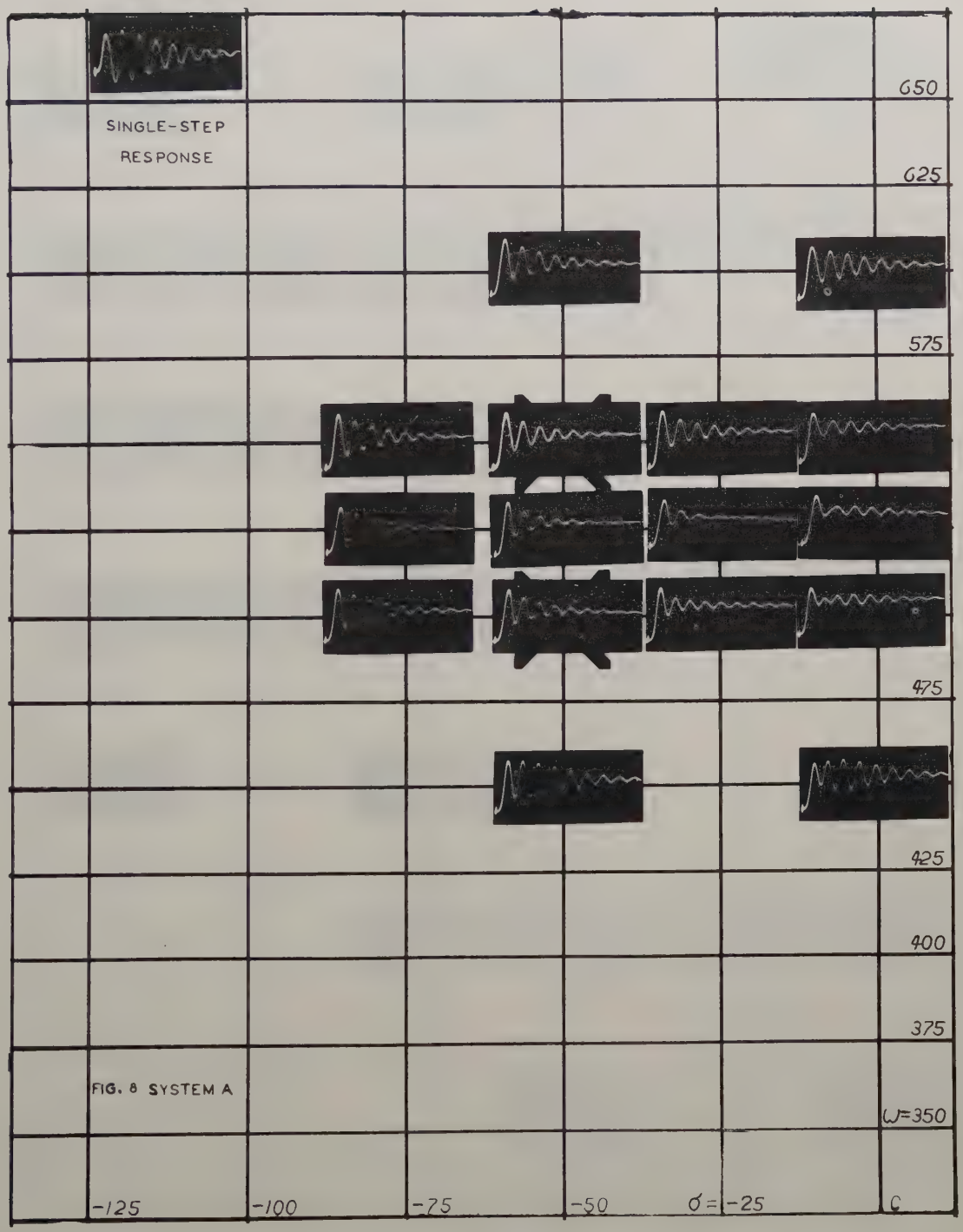


FIG. 8 SYSTEM A

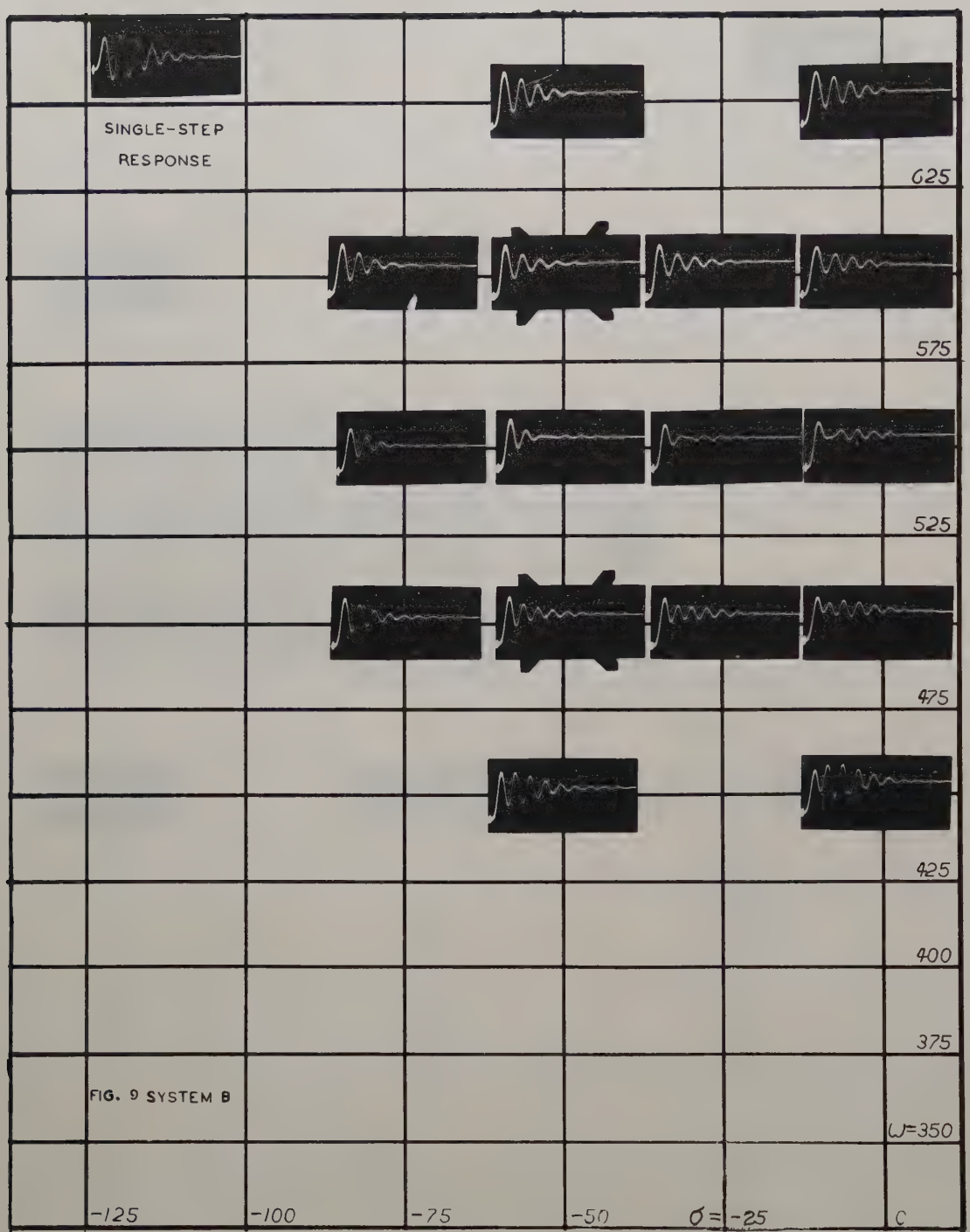


FIG. 9 SYSTEM B

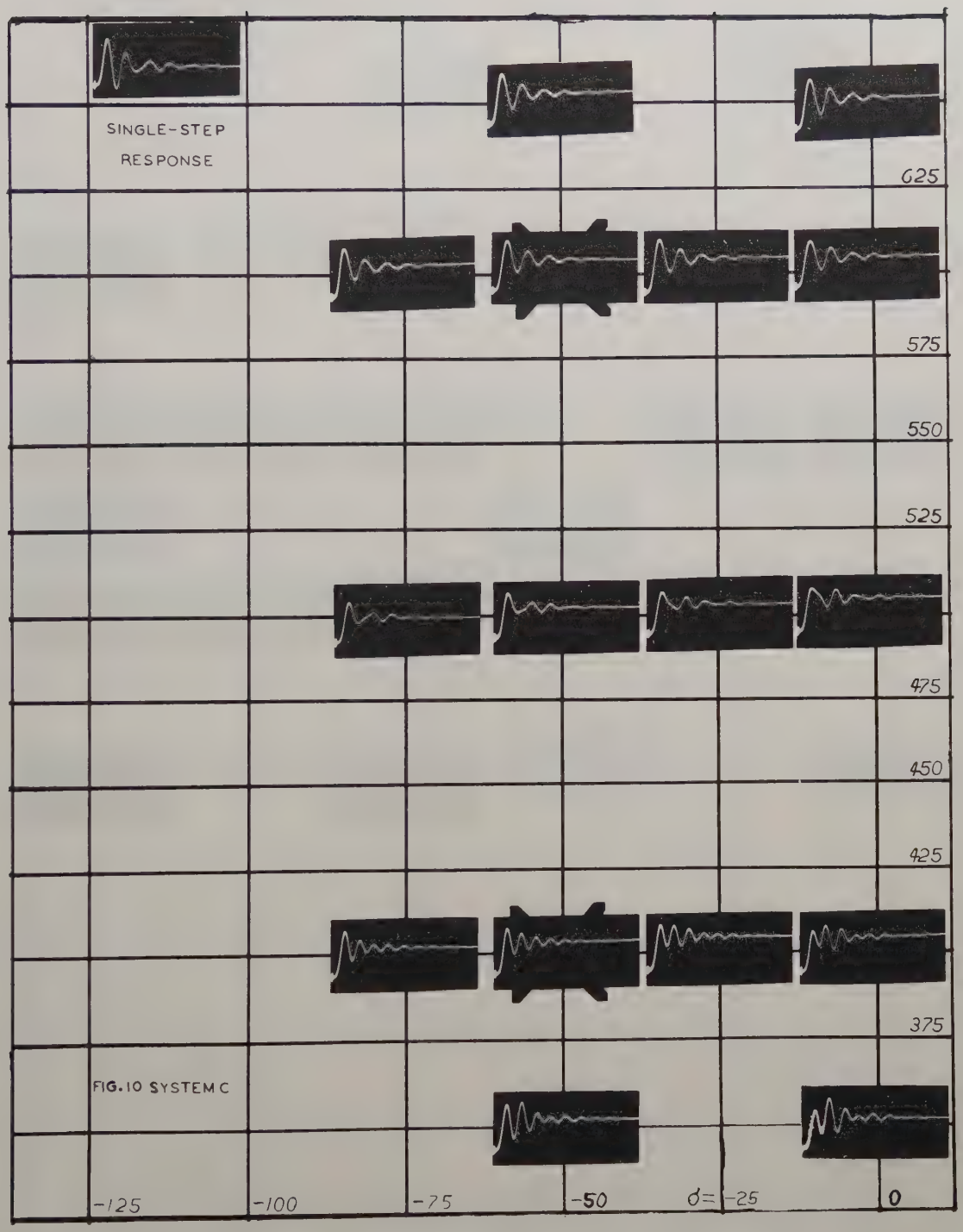


FIG. 10 SYSTEM C

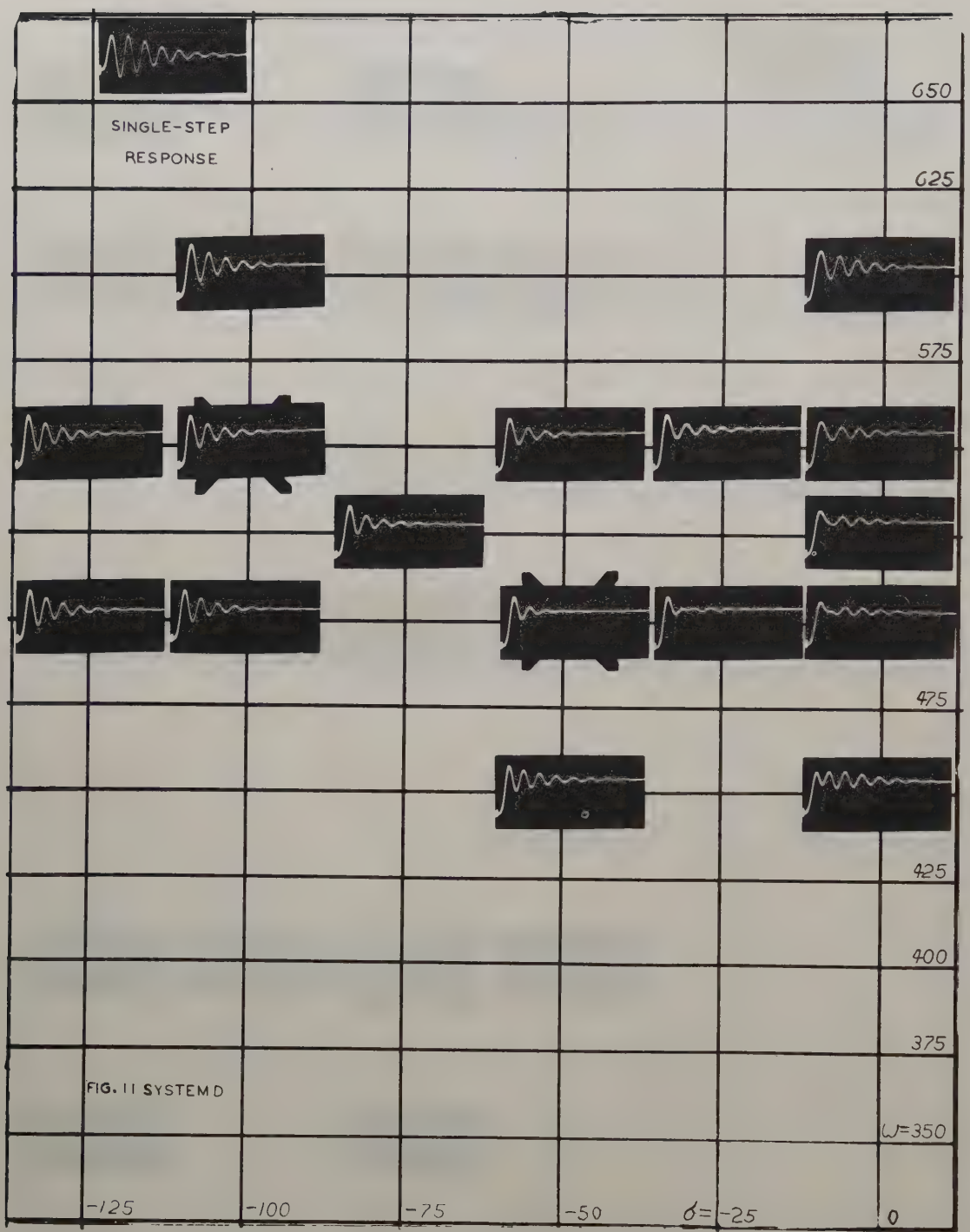


FIG. 11 SYSTEM D

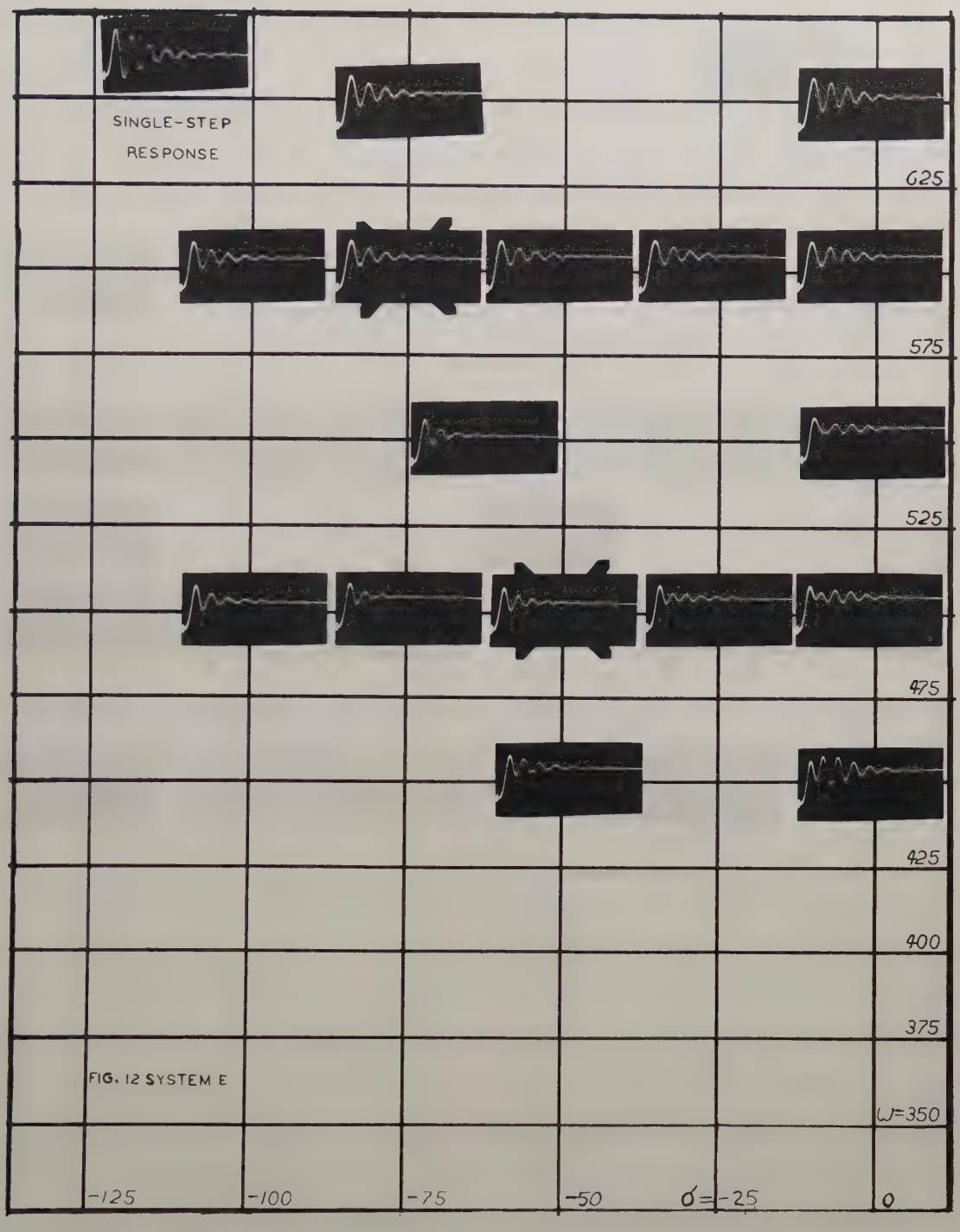


FIG. 12 SYSTEM E

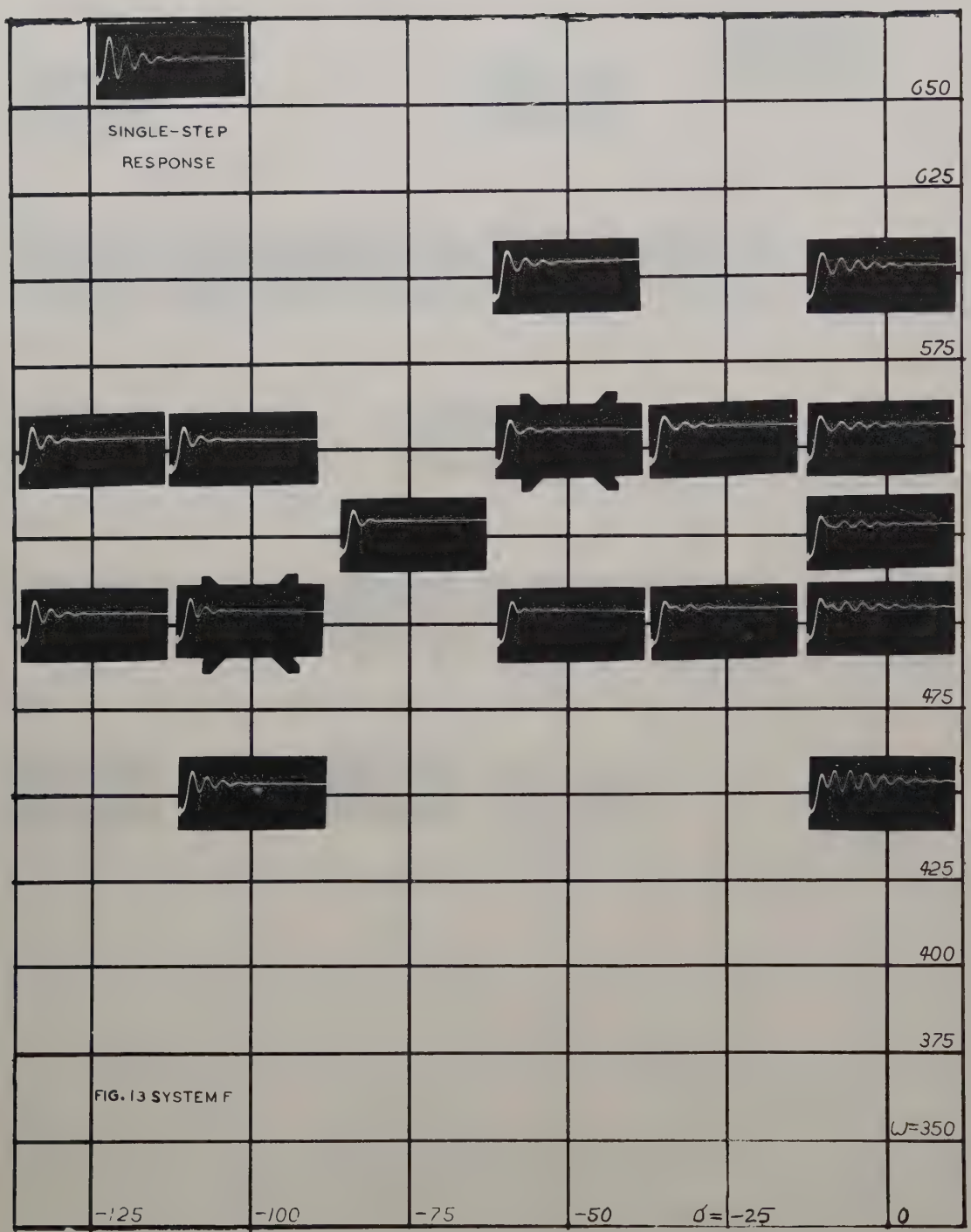


FIG. 13 SYSTEM F

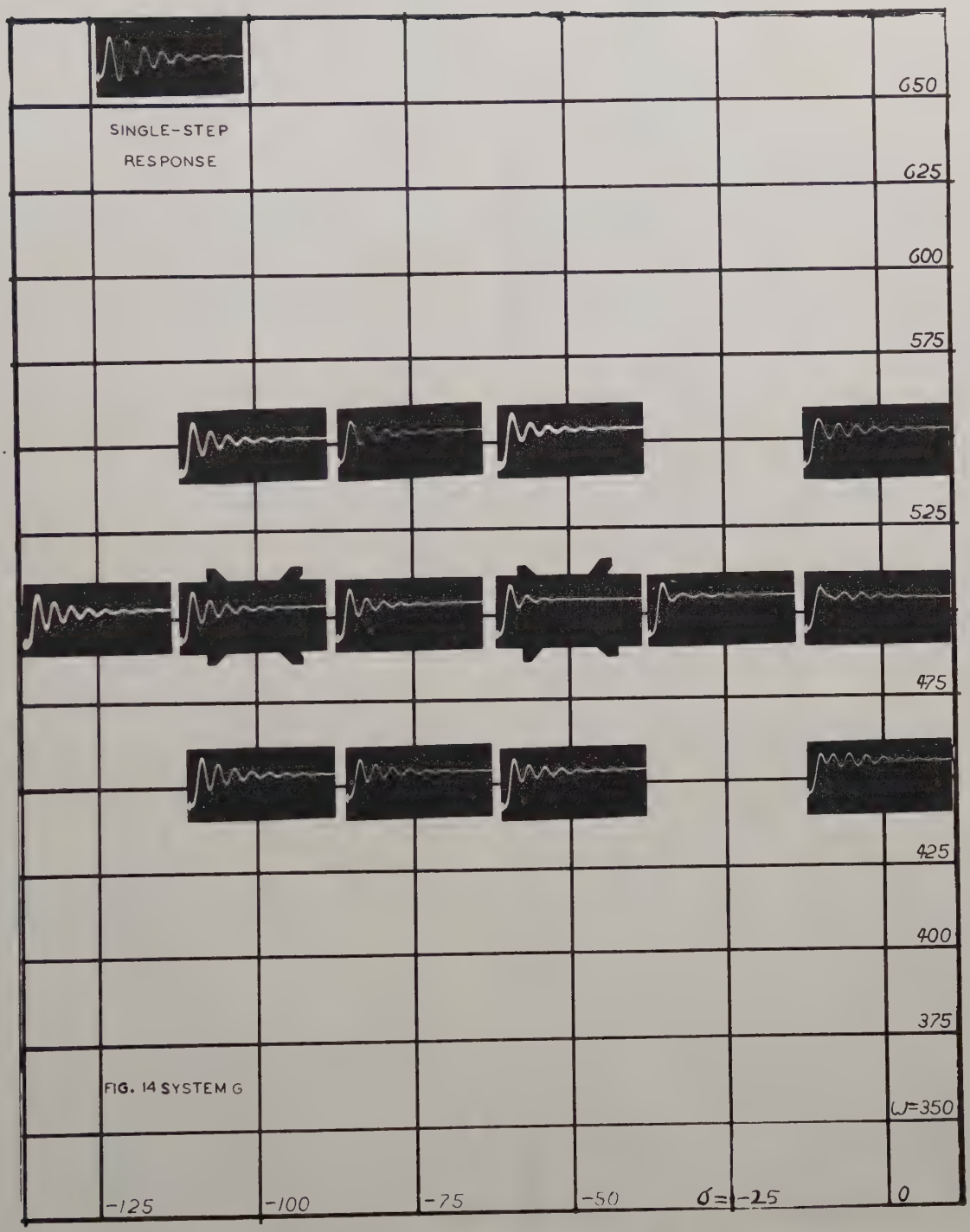
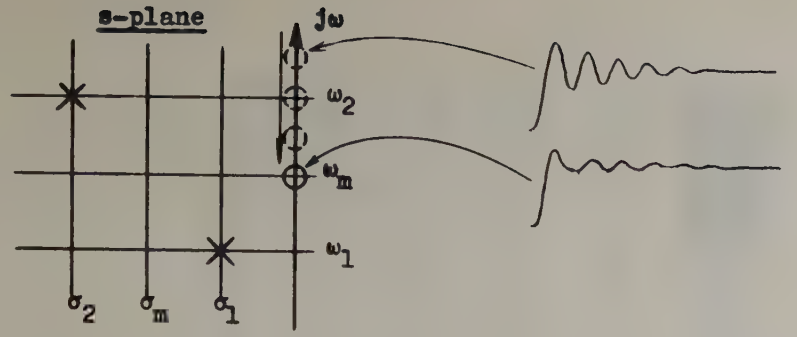
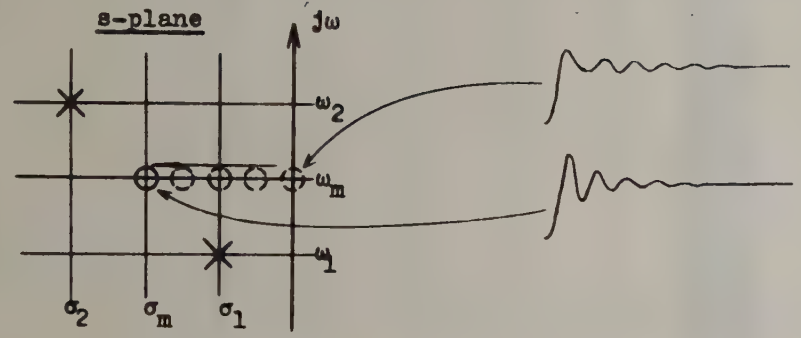


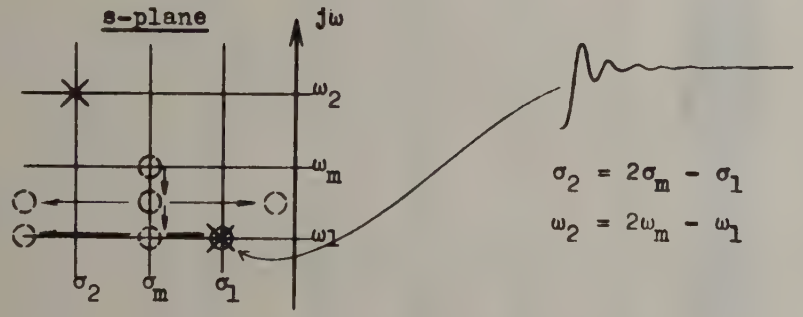
FIG. 14 SYSTEM G



(a). Step 1.

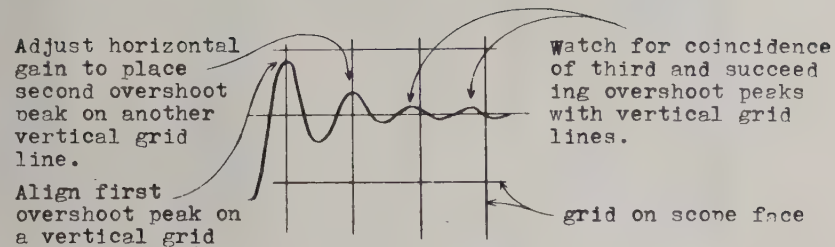


(b). Step 2.

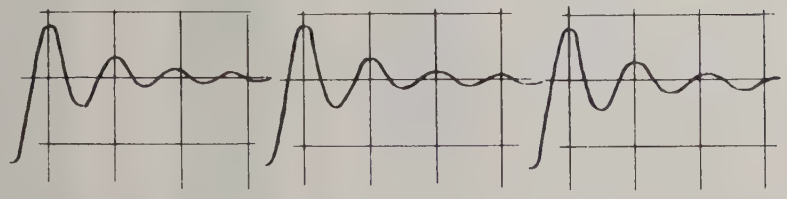


(c). Step 3.

Fig. 15. Method I System Testing



(a). Oscillograph Presentation



(b). Zero not on top of pole

(c). Zero cancels pole

(d). Zero not on top of pole

Fig. 16. Method III System Testing

RANDOM NOISE WITH BIAS SIGNALS IN NONLINEAR DEVICES

by
George S. Axelby
Westinghouse Electric Corporation

Summary

A number of investigations have been made in recent years about the transmission of Gaussian noise through nonlinear devices. In many cases, simplification or approximations were needed to make analytical solutions possible, and only zero-average Gaussian input signals were used when the results were applied to feedback control systems.

This paper presents a different approach to the problem of noise transmission through nonlinear single-valued elements. Basically, amplitudes removed by nonlinear saturation or deadzones are replaced by impulses in the amplitude distribution functions of the output signals, and the resulting first and second moments of the output distribution are computed to yield the average and rms value of the output signal. The solution may be found by graphical or mathematical integration, a visual representation of the phenomenon is obtained, and input signals with any distributions having non-zero average values may be considered.

It is shown that there is an equivalent transmission function or describing function for the average value of the noise, another for the rms value, and that one is a function of the other. Examples of the functions are given and the simpler functions with zero-average values are compared to the results obtained by other methods.

Finally, the application of the noise describing functions to feedback control systems is discussed. Theoretical results are compared with those obtained from analog simulations.

Introduction

Investigations by M. Kac,[1] A. J. F. Siegert,[2] D. Middleton,[3] J. L. Lawson and G. E. Uhlenbeck,[4] R. C. Booton,[5] and K. Chuang and L. F. Kazda[8] have provided a sound mathematical background for determining the noise output of nonlinear devices subjected to Gaussian inputs. Most of the methods developed have been concerned with the determination of probability distributions of saturated and rectified noise from which the output noise characteristics could be found. In general, these were primarily mathematical developments, and they were not particularly adaptable for analysis of control systems with nonlinear devices.

However, in 1953, Booton developed a different approach to the transmission of noise through nonlinear devices,[5] and he applied the results to control loops containing nonlinear elements.

Booton proposed to obtain an equivalent gain function of a nonlinear device by comparing the output of the nonlinearity with the output of an equivalent gain function. The equivalent gain was chosen as the one which would yield the minimum rms difference between the two outputs. The method is summarized in Figure 1 for reference. This procedure gave reasonably good correlation between measured and expected rms outputs of nonlinear elements, and it was shown how the equivalent gain, referred to as a describing function, could be used to predict the performance of a feedback control loop. However, in the process of calculating the rms output, the physical significance or meaning of the calculation was not apparent. Although physical visualization is not particularly important, or even helpful, in solving many mathematical problems, it is often useful to engineers who must develop and translate mathematical concepts into physical reality. Many times, if a physical representation of a process is available, engineers can more efficiently predict the performance of a device, perhaps find improved methods of calculation, and often obtain more insight about actual system operation through visionary processes.

Output Amplitude Distribution Functions for Certain Nonlinearities

The method of analysis described in this paper is quite different from that of Booton's. Essentially, the rms value of the output noise is found as the second moment of a simple output probability distribution function. Of course, this assumes that the noise is ergodic; that the time averages are the same as amplitude averages, and therefore time averages need not be considered in determining rms values. The output amplitude distributions are found quite easily from certain types of single valued nonlinearities which often appear in filters and control systems with or without feedback. This is done by assuming that the input amplitude distribution to the nonlinear device is known. The principle involved is shown in Figure 2. Here the input distribution is shown as a Gaussian function, although the principle is not restricted to this particular amplitude distribution. Actually, a Rayliegh exponential, or any other known distribution could be substituted for the Gaussian distribution shown.

A discussion of the saturating nonlinearity in Figure 2a will illustrate how the output distributions are obtained. From the nonlinear saturating characteristic shown in Figure 2a, it is evident that the output amplitude cannot exceed the limit levels $\pm$ L. Therefore, the probability

that the output will exceed L is zero. This indicates that the area of the output distribution from $-\infty$ to -L and from +L and $+\infty$ must be zero. However, since the nonlinear function is actually linear between -L and +L, it is assumed that the output distribution is the same as the input amplitude distribution in this region. Therefore, the probability that the output distribution will be the same as the input distribution in the central region is the area of the input distribution between the amplitudes -L and +L. This is $2A_o$ in the figure. The probability that the input amplitude will exceed these limits is equal to the areas of the input distribution between $-\infty$ to -L and the area between +L to $+\infty$. Each area is equal to $.5 - A_o$ because the total area of a probability distribution function is unity. It follows that the probability of the input exceeding the linear region is the same as the probability that the output amplitude will be exactly $\pm$L. This is represented in the output distribution by impulses of area $(.5 - A_o)$ at $\pm$L. By using this concept, the output distribution has a total area of unity while maintaining the same distribution as the input between $\pm$L with zero area outside these limits.

In Figure 2b the same reasoning also applies. Here the output amplitude can be only -E, 0, or +E. This condition is represented in the output distribution by three impulses—one at -E, another at 0, and another at +E. The impulses are assumed to have zero widths, infinite heights, and finite areas for convenience. The areas are equal to the probabilities that the amplitudes will exist at -E, 0, and +E, respectively. Of course, the probability that the output amplitude will be +E is the probability that the input amplitude will exceed +L. This is equal to the area $(.5 - A_o)$, represented by an impulse at +E in the output distribution. The same applies to the impulse at -E. The area of the impulse at the center represents the probability that the output amplitude will be zero. This is equal to the probability that the input amplitude will not exceed $\pm$L, the area of the input distribution, $2A_o$, between $\pm$L. The figures, 2c, 2d, and 2e, show other examples of probability distribution deduced from the operation of different nonlinear devices. Obviously, impulses can be used to represent the outputs of other types of nonlinearities which may be represented by discontinuous functions.

Once the form of the output distribution has been established, it becomes a relatively simple matter to compute the rms value of the output function because it is the square root of the second moment of the distribution. For the saturating nonlinearity, the problem is similar to that of finding the radius of gyration of a flat plate shaped like the input distribution with two right angle bends at $\pm$L.*

This method has the advantage of providing a physical interpretation of the mathematics involved in finding the rms output of a nonlinearity. From a pictorial representation, it is usually easier to estimate how the output will change with variation in the nonlinearity. Graphical methods may be employed with almost any degree of accuracy desired or more exact mathematical calculations can be used, and the calculations are relatively simple to make because the moments of many distribution functions have been calculated and tabulated. For example, Figure 3 is a plot of the zero, first, and second moments of a Gaussian distribution between 0 amplitude and any amplitude L. Because the moments vary with the rms value of the input distribution, the curves have been plotted in terms of σ_x/L, where σ_x is the rms value of the input.** Note that the curves give values for only half of the distribution. When plotted this way, it is possible to consider non-symmetrical functions, those with a bias value, as explained in another section. It should be noted that the zero moment, M_o, is the area of the distribution between 0 and L, the first moment, M_1, is the average amplitude, and the second moment, M_2, is the average of the amplitudes squared (the variance, the squared value, or the rms value squared) also evaluated between amplitudes 0 and L.

The second moments of some of the outputs distributions shown in Figure 2 are extremely easy to calculate. For example, the second moment of the distribution in 2c is simply $2E^2A_o$, and in the output distribution 2b, it is $2E^2(1/2 - A_o)$ or $E^2(1 - 2A_o)$. The rms value is $E\sqrt{1 - 2A_o}$. This can be evaluated by referring to Figure 3, where M_o is the same as A_o in Figure 2.

All of the output functions of Figure 2 have a zero mean value, or zero D.C. value because the distributions are symmetrical with respect to zero. This is easily verified by computing the first moment of the areas and adding them algebraically.

For example, in Figure 2b, the first moment, the D.C. value of the output, is $+E(1/2 - A_o) - E(1/2 - A_o) = 0$. It is interesting to compare this method of calculating the rms output of a nonlinear device with that proposed by Booton. The calculated results of each method are plotted in Figure 4 for a saturating component and in Figure 5 for a relay device. Although the methods are entirely different, the results are quite similar. It appears that the output values for the relay device generally vary more widely than those of the saturating device, but the average discrepancy is only about 10 percent.

To obtain a check on the calculation, a random noise generator (General Radio No. 1390A) was connected to a Zener diode which simulated a saturating device. The rms value of the saturated output was read directly as an rms voltage (using a Ballantine rms meter No. 320), and the rms value of the output voltage was divided by the rms value

*Actually, this would be the radius of gyration projected into the plane of the plate.

**The square root of the variance.

of the input. These experimental results, plotted in Figure 4, appear to agree with Booton's where the rms value of the noise is small compared with that of the saturating limits. However, at higher values of noise it apparently deviates toward the values computed from the amplitude distributions as described in this paper. Because the measuring devices were not precisely calibrated, another method was used to check the calculations experimentally.

The other method used to check the calculations was basically numerical—a digital computer (IBM 704) was programmed to compute the rms value of a random set of numbers that were subjected to a predetermined limit. This is the basic mechanism of an actual saturating device with a random input except that a relatively small sample of numbers was used to define the input instead of an infinite number which should be used theoretically. To check the number of random values that would approximate a noise input, several different numbers of values were used. In Figures 4 and 5 the noise describing function obtained from the computer using 200 and 400 values are shown.*

Of course, this is another method of calculating the noise gain or describing function, but it is difficult to obtain the insight or engineering perspective that the amplitude distribution gives, and it is not always certain that an optimum number of random values are being used.** However, the most interesting results of the digital computer calculations are that the noise describing functions obtained appear to agree very closely with those obtained from the amplitude distribution method. This is shown in Figures 4 and 5.

Non-Symmetrical Distributions

Because the moments of the output amplitude distributions in Figure 2 can be calculated rather accurately and easily when they are partially represented with impulses, it is a simple matter to extend the method to non-symmetrical distribution functions, those with a first moment, M_1, or a D.C. component in the output function. This distribution is typical of many filters and control systems where noise is superimposed on signal information which may be considered as a slowly varying, predictable bias or D.C. level. Of course, the transmission of the noise and bias signal may be calculated separately and the results combined in a linear device, but in a nonlinear component, superposition does not apply—the transmission of the D.C. component is modified by the amount of noise present and the transmission of the noise is influenced by the magnitude of the D.C. component.

The interrelations between the D.C. input, M, and noise input are represented by σX, and are illustrated in the output distribution shown in Figure 6b for a saturating element and in 6c for a relay device. Note that the impulses in these distributions are the same, but that the moment arms are different. The areas of the impulses which contribute directly to the output noise are obviously functions of the input bias, M, the type of nonlinearity involved, and the rms value of the noise, σX.

The first and second moments of the output distributions 6b and 6c can be computed from moment values obtained from Figure 3. The moments of Figure 6c may be obtained directly,† but the moments of the distribution in 6b must be obtained with respect to the zero reference of the input distribution although the moments in Figure 3 are computed with respect to a normal distribution with zero mean. However, the method of calculating the output distribution moments and the rms value of the output function from the three basic curves in Figure 3 is described in Appendix I.

The relationship between the D.C. bias and rms value of the input, the characteristics of the nonlinear device, and the D.C. bias and rms value of the output are shown more vividly in terms of equivalent gains or describing functions.

The describing function for a nonlinear element with an input consisting of a combined noise and bias signal actually consists of two functions: D_m, a bias signal transmission function, and D_σ, a noise signal transmission function. Each function is influenced by its own input as well as the input to the other. Thus, $D_m = f(M, \sigma X, N)$ and $D_\sigma = f(M, \sigma X, N)$ where M is the input bias, σX is the rms value of the input noise with respect to M, and N is the particular nonlinear characteristic. This over-all describing function is represented by the block diagram in Figure 7.

The functions, D_m and D_σ, were calculated and plotted in Figures 8 and 9 for a saturating element and in Figures 10 and 11 for an ideal relay device. The curves indicate how the signal and noise transfer ratios vary with respect to nonlinearity and the signal-to-noise ratio at the input to the nonlinear devices. A digital computer and an analog computer were used to substantiate the theoretical calculations as shown in the figures. Generally, the curves verify in a quantitative way the operation that might be deduced qualitatively.

*The number of values was increased to 600, but there was little difference between the outputs for the 600 and 400 sets.

**For example, it was found that a very large number of random values is needed to calculate a D.C. level in the output of a saturating device if it is less than .3 of the rms noise level.

† The first moment is $E(.5-A_{o1}) - E(.5-A_{o2}) = E(A_{o2}-A_{o1})$, the second moment is $E^2(.5-A_{o1}) + E^2(.5-A_{o2}) = E^2\left[1 - (A_{o1}+A_{o2})\right]$, and the rms value of the output is $\sqrt{E^2\left[1 - (A_{o1}+A_{o2})\right] - E^2(A_{o2}-A_{o1})^2}$. Numerical values for A_{o2} and A_{o1} are found from Figure 3 as shown in Appendix I.

The Saturating Describing Function

In Figure 8, the D.C. describing function, D_{ms}, through a saturating device is unity only if the D.C. input signal, M, and the rms input noise, σX, are less than .25 of the saturating level. However, as the D.C. input, M, approaches the saturating level, L, the effective signal transfer ratio, M_1/M, becomes considerably less than the expected gain of unity—if the input noise, σX, becomes equal to or greater than the saturating level. Of course, as the D.C. input level becomes greater than the saturating level, the D.C. gain becomes less than unity. However, even in this operating condition, the gain becomes progressively smaller as the input noise becomes greater because, as the input noise becomes more predominate, it becomes more evident at the saturated output. Since the noise peaks are always less than the saturating level, the average value of the output must be less than the saturating level.

Figure 9 indicates that the noise describing function, $D\sigma_S$, will be essentially unity for the saturating device until the rms value of the input noise becomes greater than .4 of the saturating level—if the D.C. input signal is essentially zero. However, as the D.C. input becomes greater, the effective noise transfer ratio, $\sigma_o/\sigma X$, becomes smaller. When the D.C. input exceeds the saturating level, the noise transfer ratio is zero for small values of noise. Then as the input noise increases $\sigma_o/\sigma X$ reaches a maximum value, although less than .5, and then it decreases slowly toward zero as the rms input noise becomes much larger than the saturating level.

The Relay Describing Function

If a small D.C. input signal is applied to the input of the relay device along with random noise, a D.C. level appears at the output.* However, as shown in Figure 10, the D.C. describing function, D_{MR}, is essentially zero if the input signal and rms noise are only .25 of the relay closing level, L. However, as the noise is increased, the D.C. level at the output becomes more pronounced, the transfer ratio increases to a maximum nearly .5, then it decreases toward zero as the noise input greatly exceeds the relay closing signal, L.

The noise describing function, $D\sigma_R$, for the relay is shown in Figure 11. It is interesting to observe that as the input D.C. level, M, becomes equal to the relay closing level, L, a very large amplification of noise occurs. Theoretically, the transfer ratio is infinity because all noise amplitudes, even those essentially zero in the input, are transformed by the relay into finite output values equal to E. However, as the rms value of the input noise increases, the transfer ratio decreases rapidly because the output noise or voltage is limited. As the D.C. input signal is increased, the relay opens, or chatters, on noise impulses creating output noise. Thus, the noise transfer ratio is essentially zero when the D.C. input level exceeds the relay closing level, $M/L > 1.0$, but as the noise level increases, the noise transfer ratio increases, reaches a maximum, then recedes toward zero as the input noise greatly exceeds the relay closing level.

Use of Noise and Signal Describing Functions in Closed Loops

The previous discussions have involved nonlinear elements that might be found in amplifiers, networks, or measuring devices which are not necessarily in feedback control loops. Nevertheless, it is possible to use the representative describing functions in a closed loop. However, because the signal and noise levels at the output of the nonlinearity depend on the signal and noise levels at the input to the nonlinearity and because both noise and signal input levels are functions of the feedback signal and the other elements in the loop, special innovations and considerations are necessary to obtain the noise and signal levels in the loop. The principles involved will be shown in a simple example and the results confirmed with experimental data, but first it is necessary to discuss two preliminary assumptions that will be made: (1) the noise input to the nonlinear device in the loop has a Gaussian distribution,** and (2) the shape of the noise frequency spectrum at the output of the nonlinear device is not greatly altered by the nonlinear element.

The assumption that the input noise has a Gaussian distribution is an engineering assumption commonly made and approximately justified when devices with relatively low bandwidths are used in the feedback control loop.† Thus it is assumed that even the rather strange and non-symmetrical amplitude distributions involving impulses in Figures 2 and 6 will be redistributed in the form of Gaussian functions by low band pass elements in the feedback loop before they reappear through the feedback path to the input of the nonlinear element.

The other engineering assumption that will be made is that the frequency spectrum of the noise will not be greatly changed in shape by the nonlinear device. There are several reasons for justifying this assumption: (a) it is assumed that the nonlinear devices have no frequency sensitive

*For small signal-to-noise ratios, this is rather difficult to detect. A long time or large sample is needed to obtain accuracy.

**Noise rectification is not considered in the example, but it could be analyzed in a similar manner.

† This is proved rigorously for certain types of random inputs in reference (8).

elements, (b) work by Middleton,* Lawson and Uhlenbeck,** and Smith[6] indicates that the frequency spectrum is not greatly changed after passing through a nonlinear element,† (c) usually the control system bandwidth is much smaller than the noise bandwidth and only the rms characteristic of the noise is significant. With these assumptions, the closed loop analysis can be made to find steady-state operating conditions.

The method of using the describing functions is outlined in Figure 12 for a saturating element in a simple type 1 position servo[7] subjected to a ramp input $X_i = X_v t$ ‡ and random noise characterized by the power density function:

$$\Phi_{ii} = \frac{2\sigma_i^2 \omega_o}{s^2 + \omega_o^2} .$$

The nonlinear element following the error signal in Figure 12 has a gain K_1, actually representing the describing function, D_{MS}, for the steady-state error, E_{ss}, and a gain, K_2, representing the describing function, D_{σ_S}, for the rms noise, σ_E, in the error signal. As shown previously, each effective gain is a function of its own input as well as the input to the other gain, and each input is a function of the closed loop. A graphical solution of the simultaneous equations involved is obtained rather indirectly by first assuming values for the steady-state error and the rms noise in the error signal and then solving for the respective ramp and noise inputs that would produce the assumed error signal and noise in closed loop operation. As shown in Figure 12, the expressions relating the closed loop errors and inputs are

$$E_{ss}\left[K_1(E_{ss}, \sigma_E)\right] = X_v$$

$$\sigma_E\left[1 + \frac{K_2(E_{ss}, \sigma_E)}{\omega_o}\right]^{1/2} = \sigma_i .$$

Each function is plotted as shown in Figure 13 against a common noise error abcissa, σ_E, for various values of E_{ss}. The steady-state error, E_{ss}, and noise error, σ_E, in the closed loop for any particular ramp input and noise input are found from the resulting graph by entering the graphs along horizontal lines at the ordinates representing the given ramp and noise inputs. These horizontal lines are extended to the right until they intersect steady-state error curves, E_{ss}, which have equal values on a common vertical line representing the noise error, σ_E. The common values of E_{ss} and σ_E found in this way are the steady-state bias error and noise error, respectively, for the given noise and signal inputs. Figure 14 shows the steady-state errors and noise errors for various ramp inputs as obtained from the graphs in Figure 13. It should be noted that if the assumed system were linear and without noise, the steady-state error, E_{ss}, would always be equal to the input. However, as the rms value of the input noise increases in the nonlinear system, the steady-state error becomes larger.

The results of this analysis were verified on an analog computer, and the experimental results are also shown in Figure 14. Figure 15 is a recording which shows how the loop variables appeared as time functions in the analog model.

It is interesting to note in Figure 13 that if X_v/L becomes unity, there is no intersection with a steady-state error curve, E_{ss}, if there is a noise input, except in the limit when E_{ss} approaches infinity. Ordinarily, without noise, the error would be unity, not infinity. Therefore, it would appear at first that the graph may be incorrect. However, this condition was simulated on the analog computer and the results are shown in Figure 16. As shown, the steady-state error grows larger, and this reduces the noise in the saturating element output, E_o. However, to maintain the desired output velocity, E_o must have a steady-state average value of exactly 1.0, but any noise which appears reduces this average value, as indicated in Figure 16. Therefore, no noise must appear at E_o, and to prevent this, the D.C. input, E_{ss}, to the nonlinearity must become larger and larger. Theoretically, noise may have infinite peaks, and therefore, theoretically, E_{ss} must become infinite to prevent noise from appearing at E_o. Of course, the D.C. level at E_o cannot exceed the saturating level which is equal to unity. As indicated in the analog computer runs, the D.C. error is growing slowly and it does not reach a steady-state value although it is 6.5 times the expected linear level. These analyses indicate that, if a servo is subjected to noisy ramp inputs, the

*Reference (3). Here Middleton shows that the frequency spectrum of wide band noise (noise with D.C. components) is not changed greatly by rectification and limiting.

**Reference (4). The authors show that a narrow band noise frequency spectrum is not altered greatly by severe clipping. The largest change of the frequency spectrum is in magnitude.

†Actually, the describing functions were developed without considering the noise time function of its frequency spectrum. This was done on the assumption that the random function was stationary and ergodic, that the amplitude averages were the same as the time averages. Of course, if the rms value of the output is different from that of the input, both the autocorrelation function and the power density function of the input must change. Here it is assumed that the output power density function changes with respect to the input density function in amplitude only so that its area is equal to the square of the rms output.

‡In a linear system the steady-state error for this input would be a constant proportional to X_v.

saturating levels must be large enough to prevent an unexpected increase in the steady-state error signal. Of course, the D.C. levels and noise in other parts of the loop may be calculated if the error signals are known.

Conclusions

It has been shown quantitatively that the transfer of a noisy signal through a nonlinear element is a function of the magnitude of the signal and noise, the signal-to-noise ratio, as well as the type of nonlinearity involved. The over-all transfer function may be considered as a special describing function with two parts—one for the noise, the other for the signal, each a function of the input to the other. This is in accordance with the fact that superposition does not apply if nonlinear devices are involved.

It has also been shown that the steady-state operation of a control loop with nonlinear elements subjected to a combined random and deterministic bias signal, or D.C. signal, can be predicted using the describing function developed.

For the special case where the bias signal or D.C. input to the nonlinearity was zero, the describing functions were in relatively close agreement with those developed by other investigators as noted. However, it is hoped that the methods used here to develop the describing functions from the amplitude distribution functions will provide another method and a deeper insight into the more general problems involved and, in so doing, perhaps permit relatively rapid, although approximate, calculations of noise and signal transmission through nonlinearities—particular for engineering purposes.

References

(1) "On the Theory of Noise in Radio Receivers with Square Law Detectors," Mark Kac and A. J. F. Siegert, Journal of Applied Physics, Vol. 18, April, 1947, pp. 383-397.

(2) "Passage of Stationary Processes through Linear and Nonlinear Devices," A. J. F. Siegert, IRE Transactions on Information Theory, Vol. 3, March, 1954, pp. 4-25.

(3) "The Response of Biased, Saturated Linear and Quadratic Rectifiers to Random Noise," David Middleton, Journal of Applied Physics, Vol. 17, October, 1946, pp. 778-801.

(4) "Threshold Signals," James L. Lawson and George E. Uhlenbeck, Vol. 24 of the MIT Radiation Laboratory Series, McGraw-Hill Book Company, Inc., 1950, pp. 56-63.

(5) "Nonlinear Control Systems with Random Inputs," Richard C. Booton, Jr., Transactions of the IRE, Professional Group on Circuit Theory, Vol. CT-1, March, 1954, No. 1, pp. 9-18.

(6) "Trends in Feedback Systems," Otto J. M. Smith, Transactions of the IRE, Professional Group on Circuit Theory, Vol. CT-1, March, 1954, No. 1, pp. 2-7.

(7) Servomechanisms and Regulating System Design, Vol. I, Harold Chestnut and Robert W. Mayer, John Wiley and Son, Inc., New York, pp. 208-212.

(8) "A Study of Nonlinear Systems with Random Inputs," Kuei Chuang and Louis F. Kazda, AIEE Applications and Industry, No. 42, May, 1959, pp. 100-105.

APPENDIX I

Graphical Evaluation of Nonlinear Output Bias and RMS Values

As suggested in the section on non-symmetrical distributions, the bias and rms values of the output of a nonlinear function can be computed from standard moment curves such as those shown in Figure 3.

A non-symmetrical input distribution is shown in Figure 6. The normal input distribution, Figure 6a, has a D.C. bias, or first moment, M, and a variance σ_X^2. The output distributions from saturating and relay devices are shown in Figures 6b and 6c. The first moments about the X = 0 reference represent the D.C. output level and the second moments represent the mean square value of the output. These moments are easily calculated as shown in the text for the relay device.

However, the first and second moments for the output of the saturating device are not as easily obtained, although the following method makes a graphical solution possible using the curves in Figure 3.

The first moment in Figure 6b may be represented by:

$$M_{10} = \int_{-L}^{+L} (M+X)\,dA + L(.5-A_{o1}) - L(.5-A_{o2})$$

$$M_{10} = M\left[A\right]_{-L}^{+L} + M_{1X} + L(.5-A_{o1}) - L(.5-A_{o2})$$

where $\left[A\right]_{-L}^{+L} = A_T = A_{o1}+A_{o2}$ (defined by the shaded areas in Figure 6a)

and $M_{1X} = M_{1X}^{+}(M < X < L) - M_{1X}^{-}(-L < X < M)$.

As shown, $A_{o1} < A_{o2}$, and normalizing:

$$\frac{M_{10}}{L} = \frac{M}{L}(A_{o1}+A_{o2}) + \frac{M_{1X}}{L} + (A_{o2}-A_{o1}).$$

The A_{o1} and A_{o2} values can be found from the M_o curve in Figure 3 by redefining the abcissa σ_X/L to be $\sigma_X/L(1 - M/L)$ when finding A_{o1} and $\sigma_X/L(1 + M/L)$ when obtaining A_{o2}.

M_{1X}^{+} and M_{1X}^{-} can be found in a similar way from the M_1/σ_X curve in Figure 3. In this case, the values found from the curve are multiplied by σ_X/L.

For example, if M/L = .25, Table I may be made from Figure 3.

The second moment for the saturating element may be calculated in a manner similar to that used to find the first moment.

$$M_{20} = \int_{-L}^{L} (M+X)^2\,dA + L^2(.5-A_{o1}) + L^2(.5-A_{o2})$$

$$M_{20} = M^2A_T + 2M\int_{-L}^{L} X\,dA + \int_{-L}^{+L} X^2\,dA + L^2 - L^2A_T$$

$$M_{20} = (M^2-L^2)A_T + 2M(M_{1X}) + M_{2X} + L^2$$

where $M_{2X} = M_{2X}^{+}(M < X < L) + M_{2X}^{-}(-L < X < M)$.

(M_{2X} can be evaluated from Figure 3 in a manner similar to that in which M_{1X} was derived.)

Normalizing:

$$\frac{M_{20}}{L^2} = \left[\left(\frac{M}{L}\right)^2 - 1\right] A_T + 2\frac{M}{L}\left(\frac{M_{1X}}{L}\right) + \frac{M_{2X}}{L^2} + 1.$$

Finally, the desired rms value, σ_o, about the output bias level is found from

$$\frac{\sigma_o^2}{L^2} = \frac{M_{20}}{L^2} - \frac{M_{10}^2}{L^2} \quad \text{and} \quad \frac{\sigma_o}{\sigma_X} = \left(\frac{\sigma_o}{L}\right)\left(\frac{L}{\sigma_X}\right).$$

LET M/L = .25: † † * * † † * **

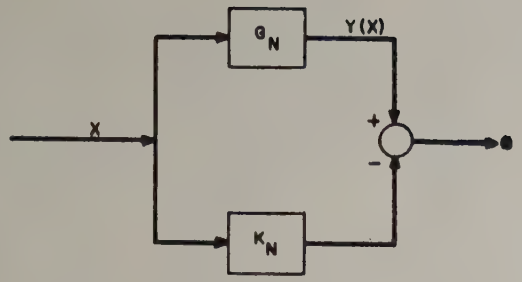
$\frac{\sigma_X}{L}$	$\frac{\sigma_X}{L(1+\frac{M}{L})}$	$\frac{\sigma_X}{L(1-\frac{M}{L})}$	A_{o2}	A_{o1}	$\frac{A_T}{(A_{o2}+A_{o1})}$	$\frac{M}{L}(A_T)$	$A_{o2}-A_{o1}$	$\frac{M_{1X}^{-}}{\sigma_X}$	$\frac{M_{1X}^{+}}{\sigma_X}$	$\frac{M_{1X}}{\sigma_X}$	$\frac{M_{1X}}{L}$	$\frac{M_{10}}{L}$	$\frac{M_{10}}{M}$
.5	.4	.667	.493	.435	.928	.232	.058	.380	.275	-.105	-.0525	.2375	.950
1.0	.8	1.333	.395	.273	.668	.167	.122	.225	.100	-.125	-.125	.164	.656

Table I

*These columns add to equal M_{10}/L.

** $\frac{M_{10}}{M} = \left(\frac{M_{10}}{L}\right)\left(\frac{L}{M}\right)$. This is the desired output-input ratio plotted in Figure 8.

†Found from Figure 3.



WHERE:

G_N IS A NONLINEAR FUNCTION GIVEN

K_N IS AN EQUIVALENT GAIN TO BE DETERMINED

ϵ IS THE DIFFERENCE BETWEEN THE ACTUAL GAIN AND THE EQUIVALENT GAIN OUTPUTS

$$\epsilon = Y(X) - K_N X$$

THE MEAN SQUARE VALUE OF $\epsilon = \overline{\epsilon^2} = \overline{Y^2(X)} - 2K_N\,\overline{XY(X)} + K_N^2\,\overline{X^2}$

AND WHEN $K_N = \dfrac{\overline{XY(X)}}{\overline{X^2}}$ $\overline{\epsilon^2}$ IS A MINIMUM.

OR $K_N = \dfrac{1}{\overline{X^2}} \displaystyle\int_{-\infty}^{\infty} X\,Y(X)\,P(X)\,dX$

Fig. 1. Booton's statistical describing function.

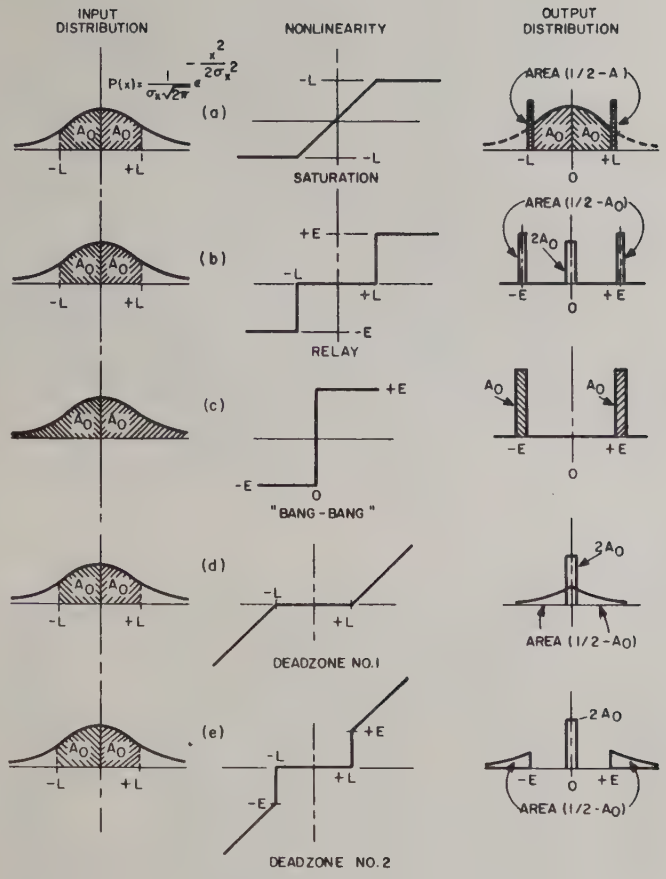


Fig. 2. Input and output distribution functions.

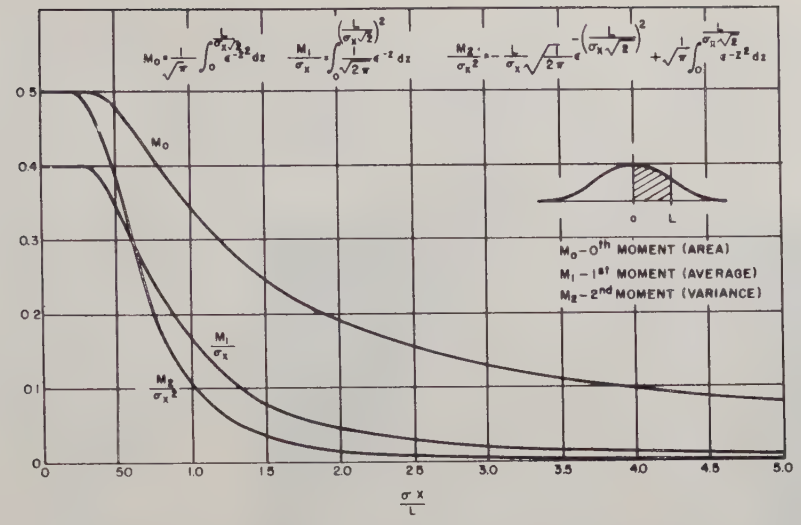


Fig. 3. Moments of a normal distribution for $0 < X < L$.

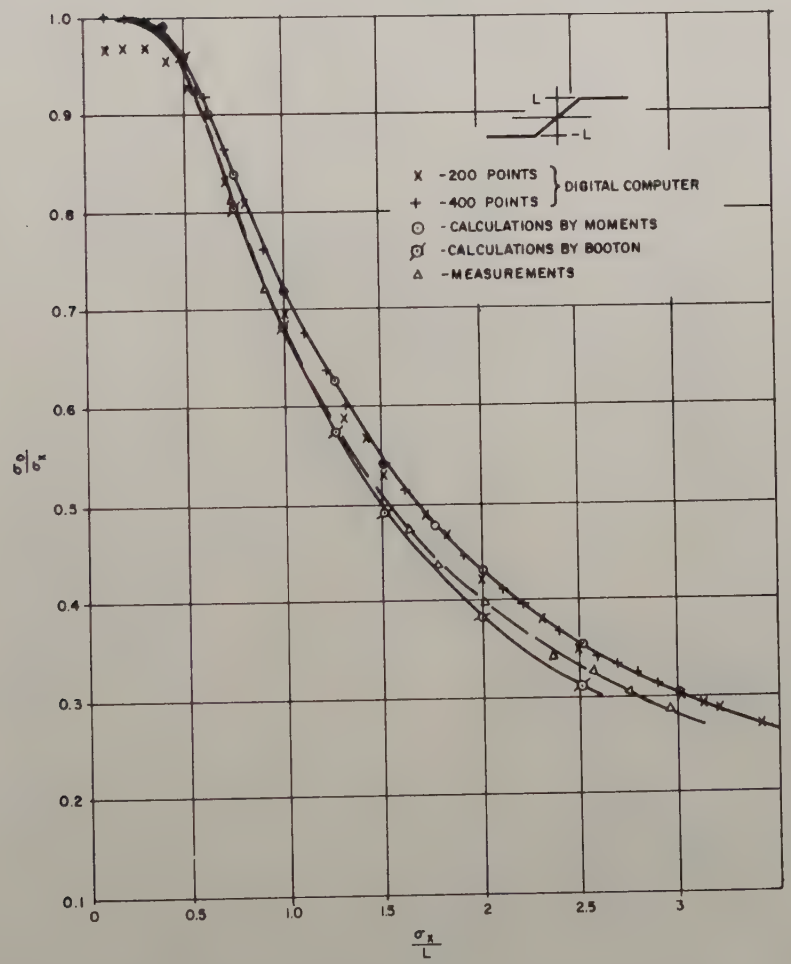


Fig. 4. Noise transmission through saturating device.

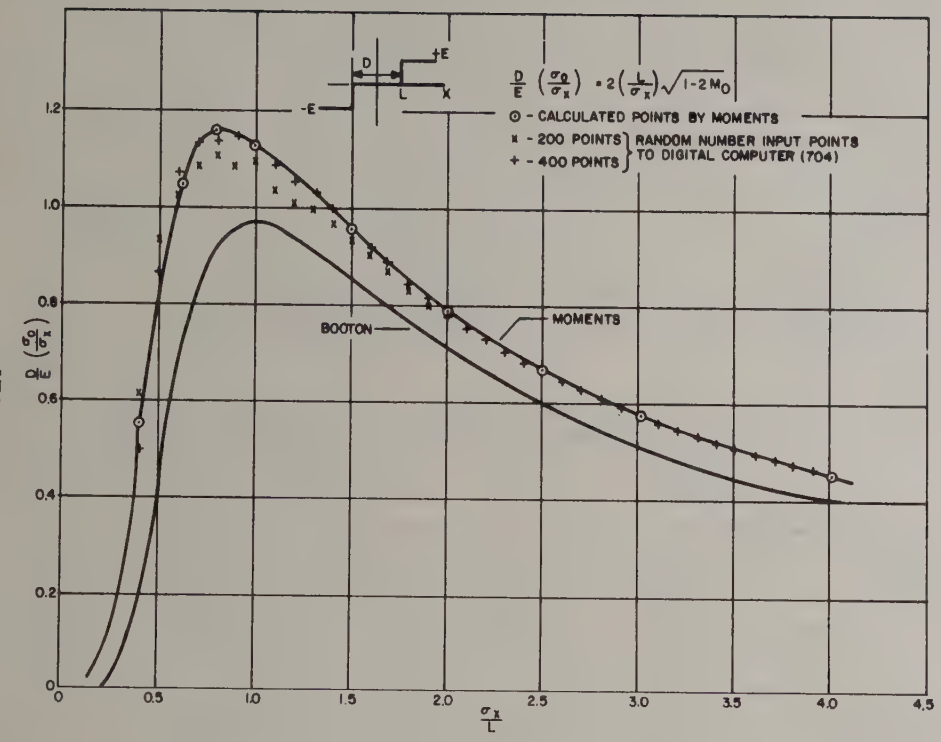


Fig. 5. Noise describing function, D_σ (no bias).

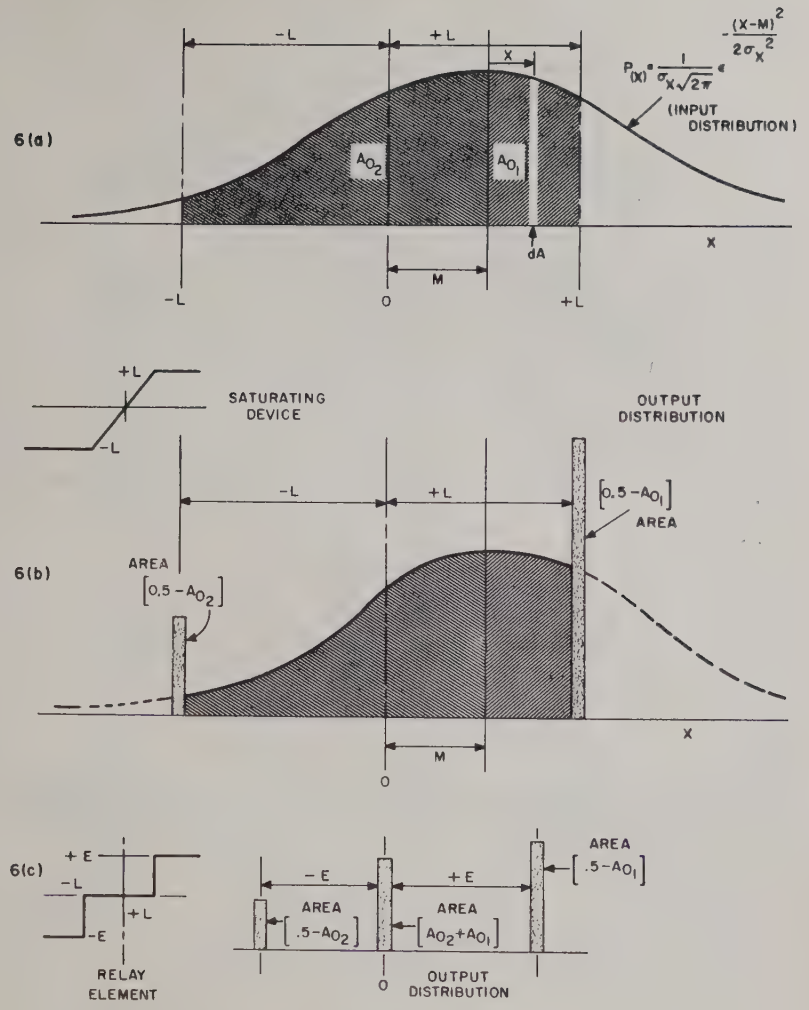


Fig. 6. Nonsymmetrical distribution functions.

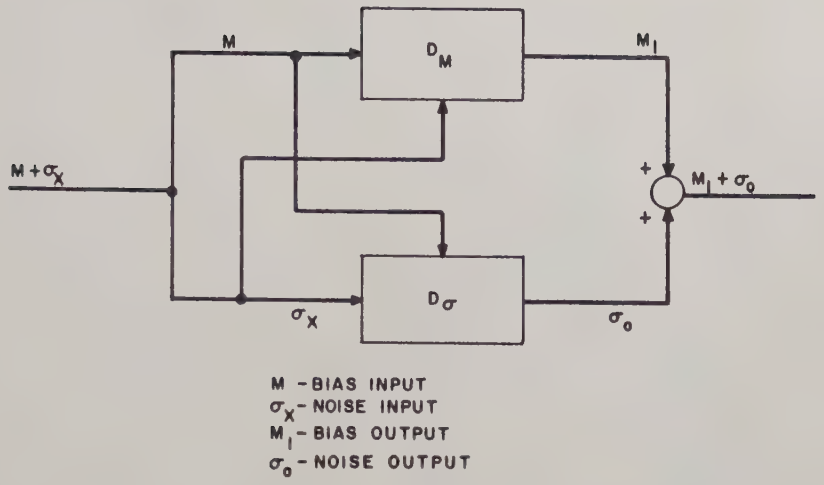


Fig. 7. Describing function for signal and noise.

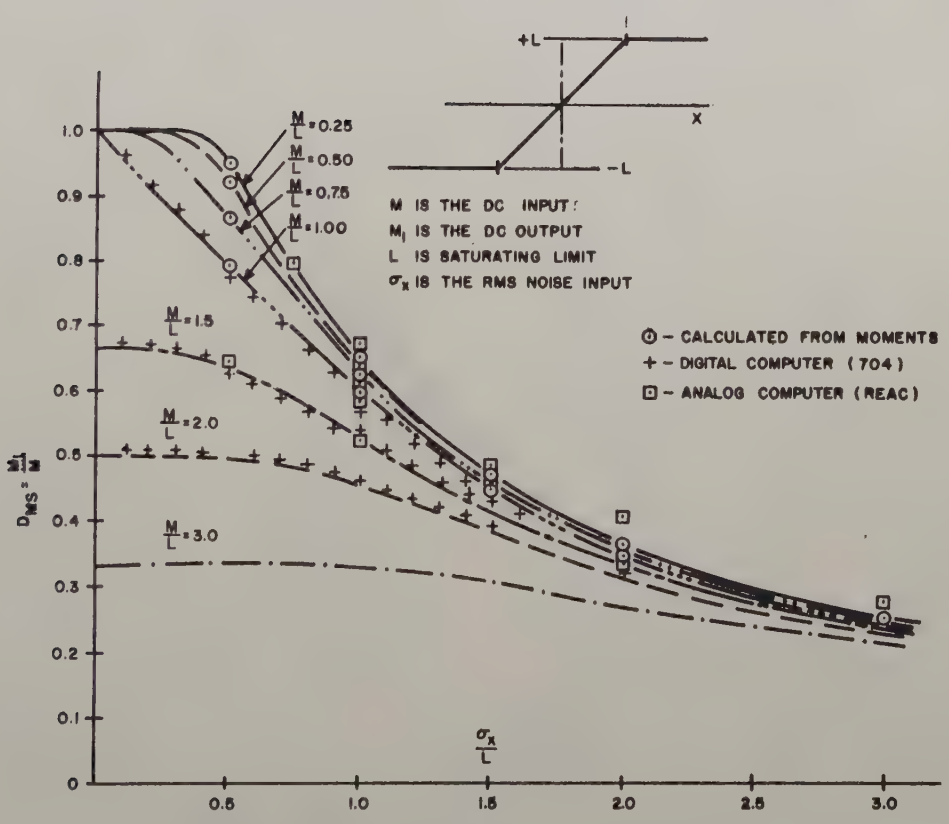


Fig. 8. DC describing function D_{MS}.

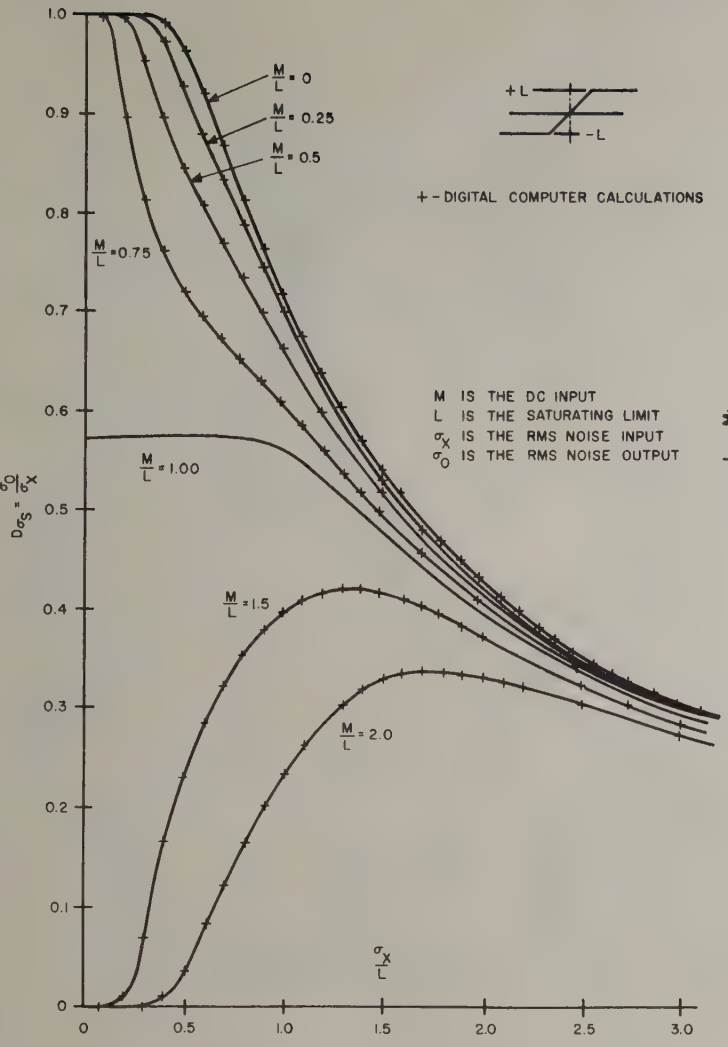


Fig. 9. Noise describing function, D_{σ_S}

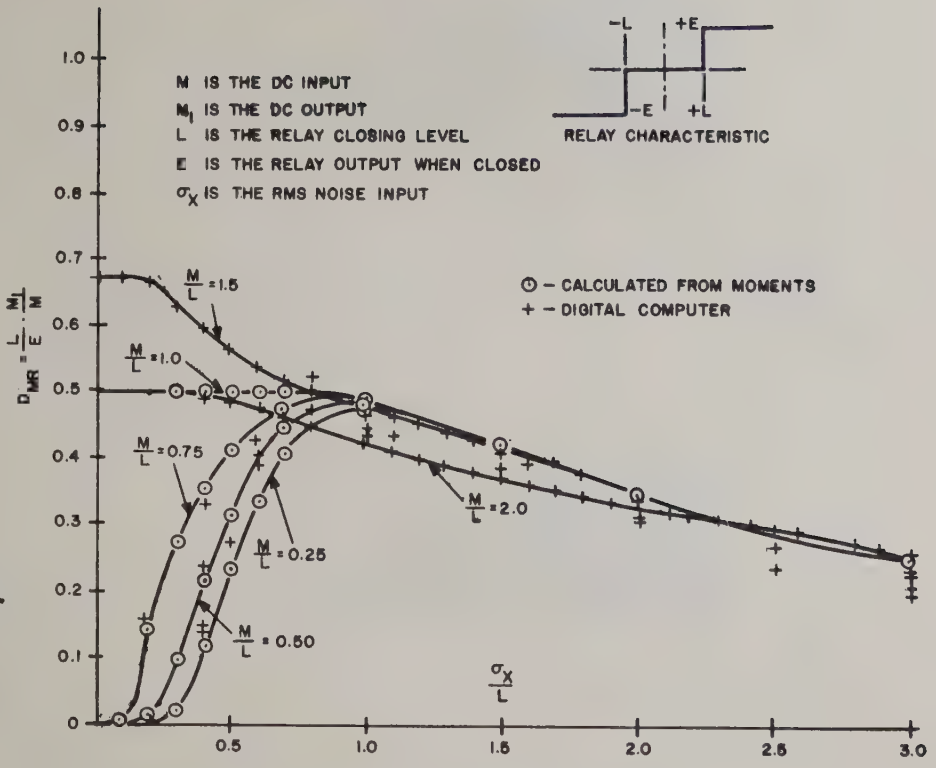


Fig. 10. DC describing function D_{MR}.

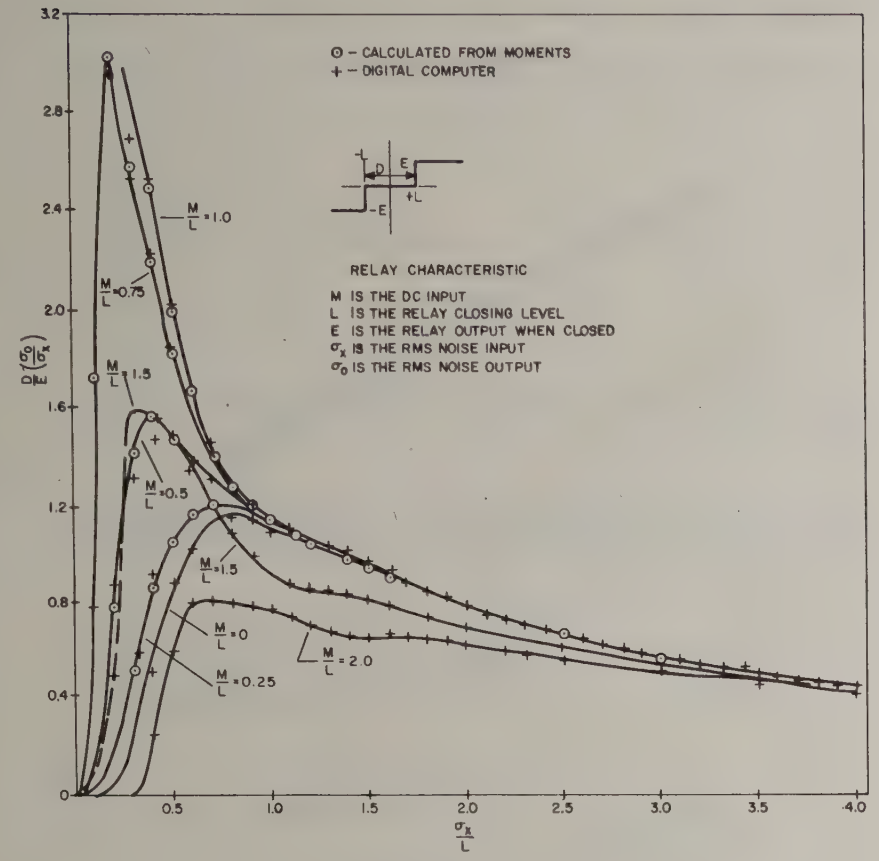
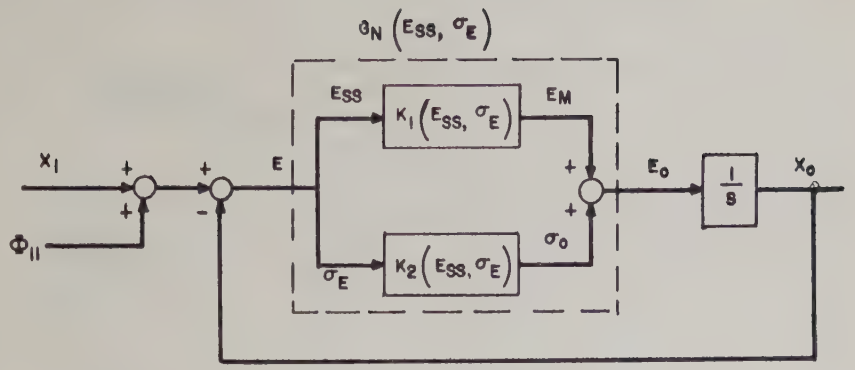


Fig. 11. Noise describing function, D_{σ_R}.



G_N IS THE NON LINEAR DEVICE

FOR RAMP INPUT $X_1 = X_v t$, $E_{SS} = \dfrac{X_v}{K_1(E_{SS}, \sigma_E)}$

FOR NOISE INPUT $\Phi_{11} = \dfrac{2\sigma_1^2 \omega_0}{\omega_0^2 - s^2}$

$$\overline{\sigma_E^2} = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty} \left(\frac{s}{s+K_2(E_{SS},\sigma_E)}\right)\left(\frac{-s}{-s+K_2(E_{SS},\sigma_E)}\right)\frac{2\sigma_1^2\omega_0}{\omega_0^2-s^2}\,ds$$

$$= \sigma_1^2 \omega_0^2 \left[\frac{1}{\omega_0 + K_2(E_{SS},\sigma_E)}\right]$$

$$\sigma_E = \sigma_1\left[1+\frac{K_2(E_{SS},\sigma_E)}{\omega_1}\right]^{-\frac{1}{2}}$$

THE INPUT FUNCTIONS MAY BE FOUND BY ASSUMING VALUES FOR E_{SS} AND σ_E IN THE FOLLOWING EXPRESSIONS:

$$E_{SS}\left[K_1(E_{SS},\sigma_E)\right] = X_v$$

$$\sigma_E\left[1+\frac{K_2(E_{SS},\sigma_E)}{\omega_0}\right]^{\frac{1}{2}} = \sigma_1$$

THESE FUNCTIONS ARE PLOTTED IN FIGURE 13

Fig. 12. Closed-loop equations.

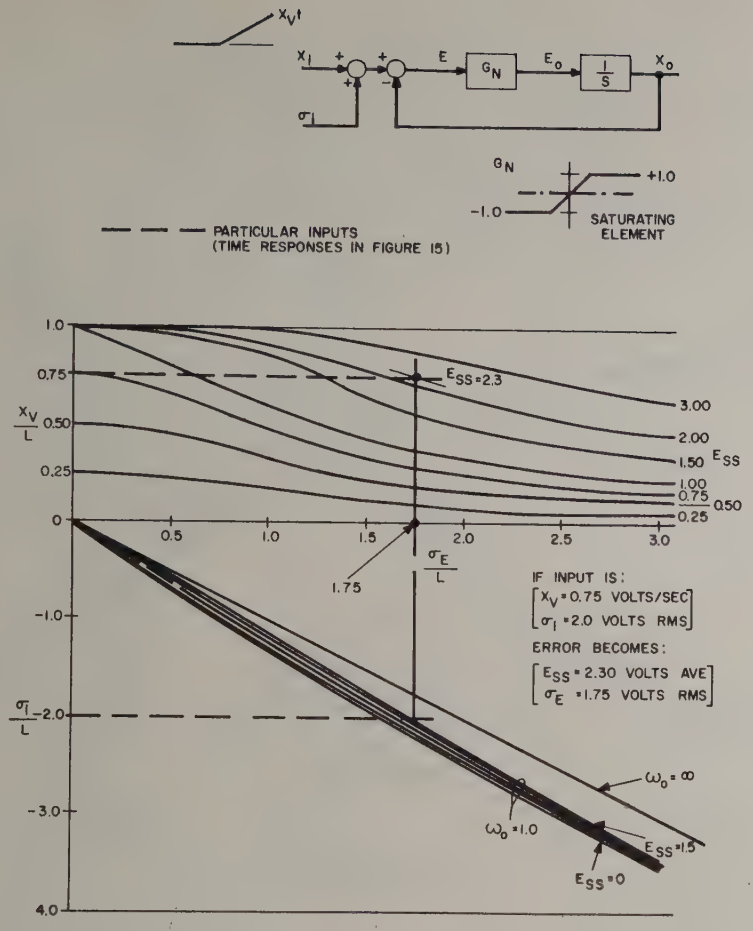


Fig. 13. Graphical method of determining steady-state closed-loop error.

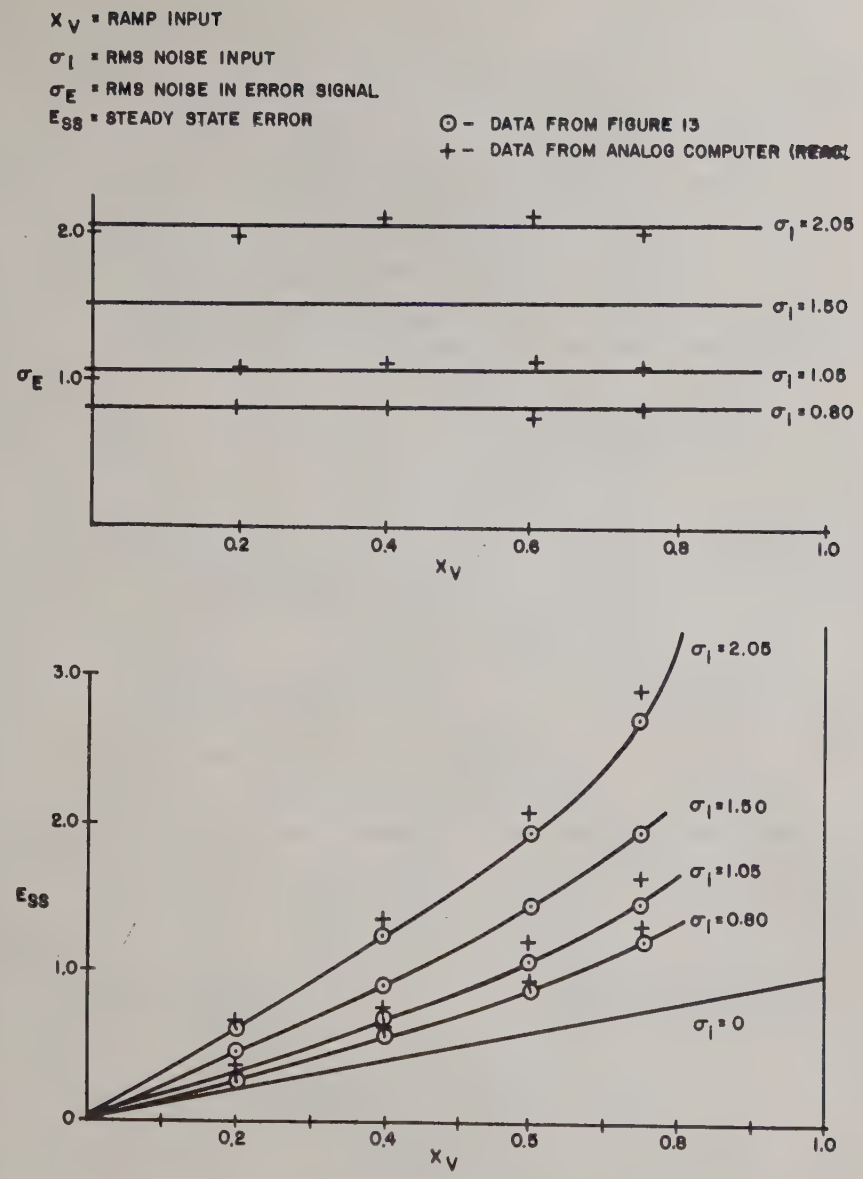


Fig. 14. Steady-state operation of closed loop.

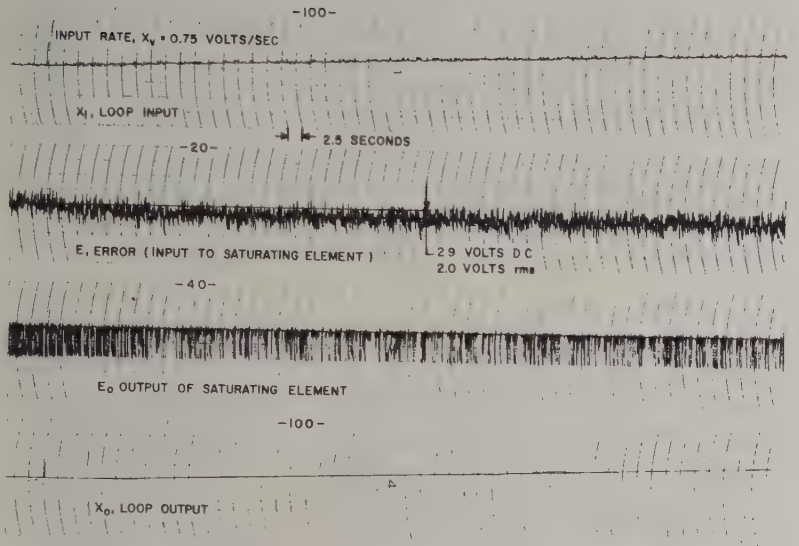


Fig. 15. Closed-loop time responses.

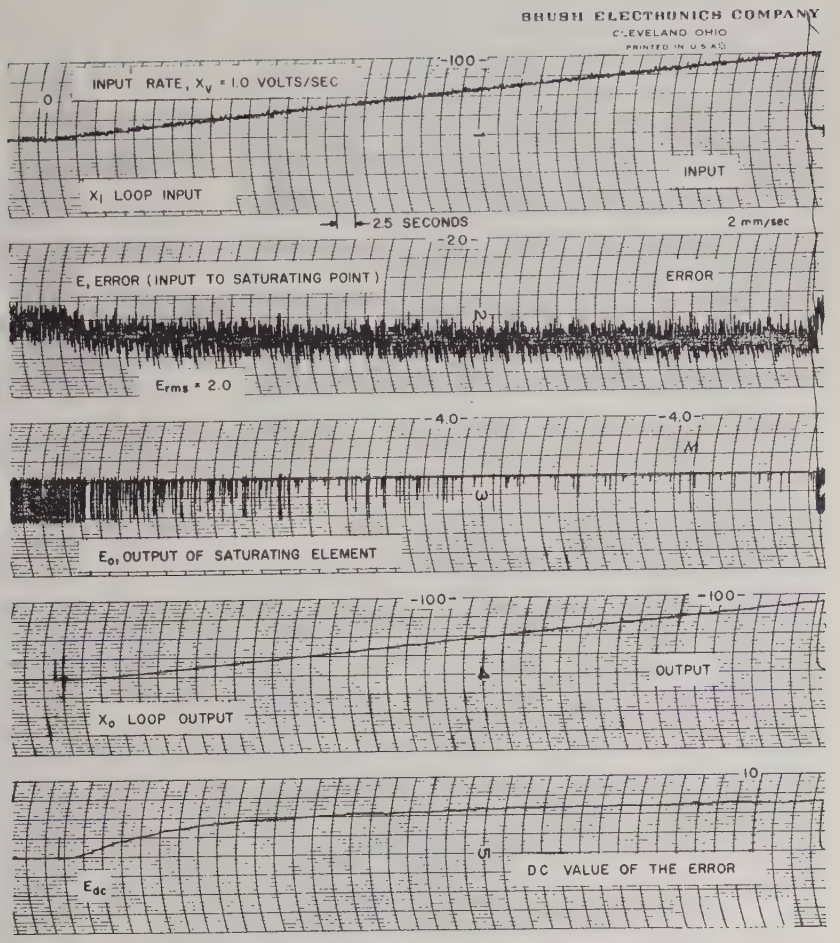


Fig. 16. Closed-loop response with error saturation.

NONGYROSCOPIC INERTIAL REFERENCE

J. J. Klein
Lockheed Aircraft Corporation
Missiles And Space Division
Sunnyvale, California

Summary

The space stabilizing capability of a servo controlled nonrotating inertial mass is analyzed and its drift rate performance compared with a high precision gyroscope. The results indicate that the nongyroscopic inertial reference and guidance system exceeds the performance of gyroscopic systems by several orders of magnitude. It is shown that a servo system materially contributes to the drift accuracy in the nongyroscopic system but not in the gyroscopic system. The navigation errors of the discussed scheme are directly proportional to the accelerometer errors and, therefore, are inherently smaller than in an equivalent gyroscopic system.

Introduction

The major part of the material presented in this paper was generated in December 1951. At that time, a significant effort was in progress at Government agencies and private industry to develop high precision gyroscopic instruments for long range inertial navigation. The initial objective of this study was to investigate ways and means to develop simple and inexpensive nongyroscopic space reference systems of low performance for target missile application, etc.* Although it soon became apparent that the potential of the applied concept was more in the field of ultra precision space reference systems, not much interest could be generated to extend these investigations in competition with the effort in the gyro field.

The presentation of this material is motivated by considerations of present efforts to develop what is commonly referred to as "exotic" gyros among which nuclear concepts appear to predominate. The reasons for this new effort originate from the requirements of future space flight navigation with emphasis on long operational life times, high accuracy and low power consumption.

Concept

The implementation of inertial guidance systems invariably calls for gyroscopic instruments of one kind or another. Newton's three laws which represent the basis for inertial guidance do not prescribe a gyro in such a system. It is, therefore, justified to investigate the possibility of implementing an inertial guidance system without the use of gyros.

In this investigation, the stabilizing capability of a nonrotating inertial mass is determined and the performance of this system is compared with a gyroscopic system. The comparison is facilitated by the postulation that no time limit of stabilization exists in either system. The procedure requires the application of servo loops around the inertial mass and the gyroscope. On the basis of the transfer functions and steady state solutions, a term by term comparison is possible and the drift rate performance can be established. Only the fundamental behavior will be discussed.

If a nonrotating inertial mass is placed in a controlled environment in such a way that the forces and moments acting on this body are predictable and consistent, it is possible to reduce their effects by servo compensation to negligible or vanishing values. In this manner, the state of rest or motion of this body in the inertial reference frame is preserved and thus can be used as an angular reference.

A possible implementation of this scheme is indicated by Figure 1. The prime element is a cylindrical body of glass which is housed in a sealed container of similar shape. The container is filled with a low viscosity fluid of the same density as that of the glass body and keeps this body in a submerged floating condition. The temperature of the whole assembly is accurately controlled in a manner equivalent to that applied to HIG gyros. The glass wheel is kept in place by thin torsion bars of quartz. The container is supported by conventional ball bearings with the axis of rotation coinciding with that of the inner wheel.

A torque free photoelectric or capacitance type of pick-off measures the relative motion between the wheel and container and supplies control signals to small compensation torque coils between wheel and case as well as to power torque motors between the case and its carrier. In addition, a linear accelerometer is mounted on the case, the output of which is used to provide a control signal to the compensation torque motors and the guidance computer if one is used. This system corresponds to a single axis inertial platform and may be expanded to a three degree of freedom device.

General Performance Functions Of Servo Controlled Inertial Wheel

At the stabilized floatation temperature,

* This study was encouraged by F. H. Andrix, Chief Engineer, Avionics Division, Bell Aircraft Corporation, Buffalo, New York

the disturbing moments acting on the glass body are: the restoring moments of the torsion bars due to the relative motion of the container and glass body, the viscous shear moments due to the relative angular velocity, and moments due to a residual pendulousness in a gravitation and/or acceleration field if a displacement of the centers of mass and buoyancy exists. The disturbing moments can be relatively easily determined by measuring the natural frequencies and the damping behavior of the glass body.

Introducing the compensation torque with gains G_V for damping, G_T for torsion and G_A for pendulousness and linear acceleration, the equation of motion of the inertial wheel is found as

$$\left\{I_W p^2 + (K-G_V)p + \left[(C-G_T)+I_W W^2_{wp}\right]\right\}\varepsilon = \left\{(K-G_V)p+\left[C-G_T)+G_A g\right]\right\}\gamma + \left(\frac{I_W}{g} W^2_{wp} - G_A\right)a \qquad (1)$$

The equation of motion of the case can be approximated by

$$I_C p^2 \gamma + M_C = M_D \qquad (2)$$

in which M_D may represent a disturbing moment due to bearing friction between case and its support. M_C is the control torque in accordance with the control equation

$$\left(g_1 p + g_0 + \frac{g_{-1}}{p}\right)(\gamma - \varepsilon) = M_C \qquad (3)$$

From equations (1), (2), (3) the motion of the case can be determined as

$$(p^5 + A_1 p^4 + A_2 p^3 + A_3 p^2 + A_4 p + A_5)\gamma = (p^2 + B_1 p + B_2)\gamma \frac{M_D}{I_C} + \frac{1}{I_C}(g_1 p^2 + g_0 p + g_{-1})\left(W^2_{wp} - \frac{G_A g}{I_W}\right)\frac{a}{g} \qquad (4)$$

Introducing the physical quantities

$$K = \frac{I_W}{T} \qquad (5)$$

$$C = I_W W^2_{wT}$$

The coefficients A_i and B_i become

$$A_1 = \frac{g_1}{I_C} + \left(\frac{1}{T} - \frac{G_V}{I_W}\right) \qquad (6)$$

$$A_2 = \left(W^2_{wT} + W^2_{wp}\right) - \frac{G_T}{I_W} + \frac{g_0}{I_C}$$

$$A_3 = \frac{g_1}{I_C}\left(W^2_{wp} - \frac{G_A g}{I_W}\right) + \frac{g_{-1}}{I_C}$$

$$A_4 = \frac{g_0}{I_C}\left(W^2_{wp} - \frac{G_A g}{I_W}\right)$$

$$A_5 = \frac{g_{-1}}{I_C}\left(W^2_{wp} - \frac{G_A g}{I_W}\right)$$

$$B_1 = \frac{1}{T_V} - \frac{G_V}{I_W}$$

$$B_2 = W^2_{wT} + W^2_{wp} - \frac{G_T}{I_W}$$

The steady state solution for a constant disturbing torque M_D on the case and a constant horizontal acceleration can be approximated by

$$\gamma_S = \frac{a}{g}\left(1 - \cos W_{wp}\sqrt{1-\frac{G_A g}{W^2_{wp}}}\, t\right) \qquad (7)$$

and the drift rate

$$\dot{\gamma}_S = \frac{a}{g} W_{wp}\sqrt{1-\frac{G_A g}{W^2_{wp}}} \cdot \sin W_{wp}\sqrt{1-\frac{G_A g}{W^2_{wp}}} \cdot t \qquad (8)$$

If the acceleration compensating gain is chosen as:

$$G_A = \frac{W^2_{wp} I_W}{g} \qquad (9)$$

and an accelerometer accuracy of

$$\frac{\Delta a}{a} = \eta_a$$

is assumed the drift rate at the end of acceleration phase is

$$\dot{\gamma}_S \approx W^2_{wp}\,\eta_a\left(\frac{V}{g}\right) \qquad (10)$$

V is the horizontal vehicle velocity gained during the constant acceleration period.

If by virtue of a very low wheel pendulousness and a highly accurate acceleration measurement a nearly perfect compensation is achieved the equation of motion, (4), reduces to

$$(p^3 + \bar{A}_1 p^2 + \bar{A}_2 p + \bar{A}_3)\gamma = (p^2 + \bar{B}_1 p + \bar{B}_2)\frac{M_D}{p I_C} \qquad (11)$$

with

$$\bar{A}_1 = \left(\frac{1}{T} - \frac{G_V}{I_W}\right) + \frac{g_1}{I_C}$$

$$\bar{A}_2 = \left(W^2_{wT} - \frac{G_T}{I_W}\right) + \frac{g_0}{I_C}$$

$$\bar{A}_3 = \frac{g_{-1}}{I_C}$$

$$\bar{B}_1 = \left(\frac{1}{T} - \frac{G_V}{I_W}\right)$$

$$\bar{B}_2 = \left(W_{\omega T}^2 - \frac{G_V}{I_W}\right)$$

The steady state solution of (11) for a constant M_D is

$$\varphi_S = \left(1 - \frac{g_0}{g_{-1}} \cdot \frac{W_{\omega T}^2 - \frac{G_T}{I_W}}{\frac{1}{\tau} - \frac{G_V}{I_W}}\right)\left(\frac{\frac{1}{\tau} - \frac{G_V}{I_O}}{g_{-1}}\right) M_D$$

$$+ \left(\frac{W_{\omega T}^2 - \frac{G_T}{I_W}}{g_{-1}}\right) M_D t \qquad (12)$$

and the steady state rate

$$\dot{\varphi}_S = \frac{\left(W_{\omega T}^2 - \frac{G_T}{I_W}\right) M_D}{g_{-1}} \qquad (13)$$

As can be seen by inspection of the coefficients $\bar{A}_i$ the dynamic behavior of the third order system is primarily determined by the gains g_1, g_o and g_{-1}. In order to obtain a reasonable estimate of the gain magnitudes involved it is postulated that the negative real root is a multiple N of the negative real part of the complex roots. In this manner the coefficients $\bar{A}_i$ and $\bar{B}_i$ may be represented in terms of the closed loop control system frequency, W_C and damping ratio, ζ. Application of this procedure yields the following relationship:

$$\bar{A}_1 \approx \frac{g_1}{I_C} \approx (2+N)\zeta W_C$$

$$\bar{A}_2 \approx \frac{g_o}{I_C} \approx (1+2N\zeta^2) W_C^2 \qquad (14)$$

$$\bar{A}_3 = \frac{g_{-1}}{I_C} = N\zeta W_C^3$$

With these expressions the steady state drift rate due to a constant friction becomes

$$\dot{\varphi}_S = \frac{W_{\omega T}^2 - \frac{G_T}{I_W}}{N\zeta W_e^3} \cdot \frac{M_D}{I_C} \qquad (15)$$

with a compensation gain of

$$G_T = W_{\omega T}^2 I_W$$

and an accuracy of compensation η_T one obtains:

$$\dot{\varphi}_S = \frac{W_{\omega T}^2 \eta_T}{N\zeta W_e^3} \cdot \frac{M_D}{I_C}$$

General Comparison With Gyro System

For the purpose of comparison a precision air bearing gyro with servo compensation similar to the scheme presented in Ref. 1 is chosen. According to Fig. 2 the equations of motion of the measuring (φ) axis and precession (α) axis are:

$$I_\varphi p^2 \varphi + H p \alpha + M_C = M_{D\varphi} \qquad (16)$$

$$(I_\alpha p^2 + K_a p)\alpha - H p \varphi = M_{D\alpha} \qquad (17)$$

The control torque M_C on the φ axis is obtained from the precession angle α and is chosen as

$$(a_2 p^2 + a_1 p + a_o)\alpha = M_C \qquad (18)$$

The system equation of motion is then

$$\left(p^3 + \bar{\bar{A}}_1 p^2 + \bar{\bar{A}}_2 p + \bar{\bar{A}}_3\right)\varphi \qquad (19)$$

$$= \left(p + \bar{\bar{B}}_1\right)\frac{M_{D\varphi}}{I_\varphi} + \frac{1}{p}\,\frac{M_{D\alpha} a_o}{I_\alpha I_\varphi} + \frac{M_{D\alpha} a_1}{I_\alpha I_\varphi}$$

with

$$\bar{\bar{A}}_1 = \frac{K_a}{I_\alpha} + \frac{H}{I_\alpha I_\varphi}\, a_2$$

$$\bar{\bar{A}}_2 = \frac{H^2}{I_\alpha I_\varphi} + \frac{H a_1}{I_\alpha I_\varphi}$$

$$\bar{\bar{A}}_3 = \frac{H a_o}{I_\alpha I_\varphi}$$

$$\overline{\overline{B_1}} = \frac{K_a}{I_\alpha}$$

$$H = I_W \cdot \Omega$$

The steady state solution for constant disturbing torques such as dry friction $M_{D\varphi}$ and a mass unbalance or turbine torque $M_{D\alpha}$ is:

$$\varphi_S = \left[1-(H+a_1)\frac{M_{D\alpha}}{M_{D\varphi}}\right]\left(\frac{M_{D\varphi}K_a}{I_\varphi H a_0} + \frac{M_{D\alpha}a_1}{I_\alpha I_\varphi X H a_0}\right) + \frac{M_{D\alpha}t}{H} \qquad (20)$$

$$\dot{\varphi}_S = \frac{M_{D\alpha}}{H} \qquad (21)$$

In order to facilitate the performance comparison of the compensated and servo controlled inertia wheel with a high performance air bearing gyro, the physical system parameters yielding the same drift rates in either system are compared with each other. Thus the drift rate due to an air bearing turbine torques $M_{D\alpha T}$ as compared with the drift rate of the inertia wheel system due to incomplete torsion bar compensation but perfect acceleration compensation is established as

$$\frac{1}{\Omega}\frac{M_{D\alpha T}}{I_{WG}} = \frac{W_{WT}^2 \eta_T}{N\mathcal{J} W_C^3}\frac{M_{D\varphi}}{I_C} \qquad (22)$$

If the ratio of disturbing torques to inertias Z is introduced such that

$$Z = \frac{M_{D\alpha T}}{M_{D\varphi}}\frac{I_C}{I_{WG}} \qquad (23)$$

the following relation between gyro rotor speed Ω and the servo parameters W_C, N and $\mathcal{J}$ is obtained

$$\Omega = Z\frac{N\mathcal{J}W_C^3}{W^2_{WT}\eta_T} \qquad (24)$$

A conservative assumption of the quantities is as follows:

$N = 10$ $\quad \frac{M_{D\varphi}}{I_C} = .1$

$\mathcal{J} = .07$ $\quad \frac{M_{D\alpha}}{I_{WG}} = 10^{-6}$

$W_C = 2000$ $\quad Z = 10^{-5}$

$\eta_T = .01$

$W_{WT} = .1$

With these values an air bearing gyro yielding the same drift rate as the servo compensated inertia wheel would require a rotor speed of the gyro of

$$\Omega \approx 5.6 \times 10^9$$

$$\approx 5.6 \times 10^{10} \text{ RPM}$$

Similarly, the drift rate due to an unbalance torque on the gyro precession axis as compared with the drift rate of the inertia wheel system due to incomplete pendulousness compensation can be established. The gyro unbalance torque considered here is caused by a center of gravity shift of the rotor due to nonideal bearings and assymetrical geometry changes due to heat flow from the gyro motor. The unbalance torque of the gyro in a gravitation and/or acceleration field may be expressed as

$$M_{D\alpha P} = W^2_{\alpha P} I_\alpha \frac{a}{g} \qquad (25)$$

so that the following relation exists between the two systems:

$$\frac{W^2_{\alpha P} I_\alpha \frac{a}{g}}{\Omega I_{WG}} \approx W^2_{WP}\eta a \frac{V}{g} \qquad (26)$$

$$\Omega = \frac{W^2_{\alpha P}}{W^2_{WP}}\frac{I_\alpha}{I_{WG}}\frac{a}{V}\frac{1}{\eta a}$$

Since the drift rate of the servo controlled inertia wheel system is proportional to the velocity gained during an acceleration phase, the comparison requires a specification of acceleration and velocity. With the following assumptions:

$$\frac{I_\alpha}{I_{WG}} = 5 \qquad \frac{a}{V} = 10^{-2}\ s^{-1} \qquad \eta_a = 10^{-5}$$

$$\omega_{\alpha P} = 5 \times 10^{-3}\ s^{-1}$$

$$\omega_{WP} = 10^{-3}\ s^{-1}$$

the required gyro rotor speed yielding the same drift rate as the inertia wheel system would be

$$\Omega \approx 1.25 \times 10^5 \approx 1.25 \times 10^6\ rpm$$

Accuracy Limitations

Disturbances acting on the glass body that cannot be compensated are those caused by Brownian motion and temperature field gradients, which lead to rotational motion of the floatation medium and eventually to a drift rate of the inertial wheel system. The random drift uncertainty due to Brownian motion can be estimated from

$$\Delta\bar{\varphi} \approx \sqrt{\frac{KT}{\omega_{WT}^2 \eta_T I_W}} \approx 10^{-3}\ deg. \qquad (27)$$

where K is the Boltzmann constant and T is the absolute temperature. The exact determination of the flow field in the floatation fluid due to temperature gradients is beyond the scope of this paper. However, on the basis of the fundamental behavior of a viscous fluid with density variations in a gravitational field as derived from the Navier-Stokes equations in connection with heat transport and heat conduction the tangential flow velocity U on the wheel can be derived as

$$U = \frac{\Delta T}{T} \sqrt{g \bar{H} \frac{\bar{a}}{\eta}}\ e^{-\frac{r}{L}} \sin\frac{r}{L} \qquad (28)$$

in which

$$\bar{H} = \frac{2T}{\Delta T r_o} \qquad L = \sqrt[4]{\frac{4 \eta \bar{a} \bar{H}}{g}}$$

with

$$r_o = 6\ cm$$

$$r = 3\ cm$$

$$\Delta T = .01\ deg.\ C$$

$$T = 300\ deg.\ Kelvin$$

one obtains

$$U \approx 10^{-9}\ cm\ s^{-1}$$

and

$$\Delta\dot{\varphi} \approx 10^{-4}\ deg.\ hr^{-1}$$

The integral over the statistically distributed velocities around the glass body, however, will yield a much lower net value. A velocity profile is shown in Fig. 3. Inaccuracies in the temperature level control affect the pendulousness of the wheel in the following manner:

$$\frac{\Delta \omega_{WP}^2}{\omega_{WP}^2} = \frac{\frac{\partial \varrho_F}{\partial T} - \frac{\partial \varrho_\omega}{\partial T}}{\varrho_F} \qquad (29)$$

Since, however, the acceleration compensation gain G_A can be adjusted automatically to the actual fluid temperature by means of a multiplier resistor in the fluid a better than 95% compensation of this effect which in itself is small can be achieved.

Applicable Technology

It is evident from the discussion of the various error sources that the wheel should have a high dimensional stability and accuracy as well as a high degree of homogeneity. A cylinder of glass of optical quality appears to fulfill these requirements very well. Also, the methods of high precision shaping are well established in the optical industry. In addition, glass is corrosion resistant and a great variety of densities can be achieved. The requirements on linearity and hysteresis of the torsion wire suspension can best be fulfilled by quartz which is extensively used in high precision measurements. As floatation fluid methylene bromide ($CH_2 Br_2$) having a density of 2.46 gcm^{-3} at 30 deg C may be used. It has a melting point of -52.8 deg C and a boiling point of 97.8 deg C. The viscosity is low and amounts to 1.225 centipoise at 25 deg C. If the wheel configuration has a moment of inertia of 1 cmg s^2 and a spacing of 3 cm is chosen the time constant will amount to approximately 300 seconds and no compensation for viscosity is required (G_V=0).

A relatively crude experimental two metal combination (magnesium and copper) wheel verified the obtainable time constants. The pendulousness was established as $W_{wp}=3\times10^{-3}\ s^{-1}$ inspite of corrosion difficulties. It is expected that a good glass wheel will have value of $W_{wp}=10^{-4}\ s^{-1}$ or better. Capacitance type of transducers or optical pickoffs may be used to furnish the compensation and control signals. Since the torque level of the compensation is very low galvanometer type of torquers with a few windings are applicable.

The servo motor between case and vehicle frame should not introduce disturbances due to airframe motions. For this reason, gears should be avoided and direct torquers be used. DC torquers with permanent magnet excitation are favored over other schemes for reasons of overall efficiency. Since solid state devices such as transistors or controlled rectifiers should be used in the power amplifier a pulse modulation scheme is preferable. A possible scheme is shown in Fig. 4. This method makes use of a polarized relay as tne modulation element but equivalent circuitry can be developed which would result in a similar operation. The relations between signal flux ϕ_s, feedback flux ϕ_o and dead zone flux ϕ_D are shown in Fig. 5. A first order analysis yields an "on" time of

$$T_1 = \frac{2\tau\phi_D}{\phi_o}\,\frac{1}{1-P} \qquad (30)$$

where P is the ratio between signal and feedback flux the "off" time is

$$T_2 = \frac{2\tau\phi_D}{\phi_o}\,\frac{1}{P} \qquad (31)$$

The frequency of the modulation is then

$$f_M = \frac{\phi_o}{2\phi_D\tau}\,(P-P^2) \qquad (32)$$

so that the modulation is

$$M = f_M T_1 = P \qquad (33)$$

With this the effective control voltage V accross the torque motor in relation to the power supply voltage V_o is obtained as

$$V = \frac{V_o}{\phi_o}\,\phi_s \qquad (34)$$

These relations are plotted in Fig. 5.

Potential Application

Terrestial Navigation

The servo controlled inertia wheel system is ideally suited for long range terrestial navigation if it is operated in the Schuler tuning mode. The equation of motion of the angular reference which indicates the local vertical undisturbed by vehicle acceleration is similar to (4). The acceleration compensating gain has to be adjusted to yield a natural frequency due to the gravitational field of

$$W_s^2 = \frac{g}{R} \qquad (35)$$

where R is the earth radius. Double integration of the accelerometer output yields the distance traveled. The distance errors due to incomplete compensation of the torsion wire suspension and fluid viscosity are very small and may be neglected entirely. The predominant term of the distance error due to incomplete acceleration compensation is derived as

$$\Delta s = \frac{W_{wp}}{W_s^2}\,\eta_a s \qquad (36)$$

Assuming the same physical system constants as before, the distance error after a 6,000 mile travel at a speed of 1,000 ft. s^{-1} amounts to 200 ft. The lateral deviations from the intended course originate from the pendulousness of the inertial wheel used for azimuth control, accelerometer errors and the bearing friction of the case. The deviations due to the first two error sources are predominant and amount to

$$\Delta y_P = \frac{W_{wp}^2}{2g}\,\eta_a s^2 \qquad (37)$$

The lateral deviation due to bearing friction on the case is

$$\Delta y_{MD} = \frac{W_T^2}{2Nf W_c^3}\,\eta_T\,\frac{M_D}{I_c}\,\frac{s^2}{V} \qquad (38)$$

With the previously given values the total lateral deviation after 6,000 miles amounts to approximately 200 ft.

Space Operation

The servo controlled inertia wheel system may also be applied advantageously to satellite and space vehicle control and guidance. The electrical power consumption is much lower than in gyroscopic devices with high speed motors. The operational life time depends only on the mean time to failure of the associated electronic components and thus can be expected to amount to years and decades. In addition, this system is much better adapted to the hostile environment of an orbital boost phase such as high acceleration and vibration levels. The drift rates in space will be lower than in terrestial applications since the weightlessness eliminates all errors due to pendulousness and density variations. Thus, the drift rates are only due to bearing friction, incomplete torsion wire compensation and Brownian motion. Under these conditions the drift rates can be expected to be of the order of 10^{-9} deg/hr.

Initial Conditions

The nongyroscopic inertial reference system needs in contrast to gyro systems intelligence as to initial rate conditions. In terrestial applications this requirement is fulfilled by caging the device at the launch or take off site. In this manner the appropriate earth rate component is automatically introduced. In other applications where the requirement exists to generate an angular rate equivalent to gyro precession this characteristic may be even advantageous since an initial impulse to the compensation torque is sufficient to accomplish this. In satellite and space vehicle operations, the desirability exists for correcting the drift rates acquired in the boost phase by monitoring with tracking devices such as star trackers or sun followers. Since these optical or infrared devices can achieve accuracies of fractional seconds of arc, it appears that the accuracy potential of this device can be put to practical use in space navigation also.

Conclusion

The analysis of the nongyroscopic inertial reference and guidance system shows that the application of conventional servo techniques in combination with well established technologies leads to an accuracy potential that cannot be achieved by gyroscopic instruments. The basic reason for this is that the servo system in the case of the described system materially contributes to an accuracy improvement whereas the servo system in the case of gyros does not. Also, the system accuracy in the nongyroscopic navigation system is directly proportional to the accelerometer accuracy which is not the case for gyroscopic navigation systems. This fact is of high significance since experience has taught that it is easier to build highly accurate accelerometers than to build highly accurate gyros.

Ref. 1

The Air-bearing Gyro Stabilization Problem
John P. Jagy, Missiles and Rockets, February 1958

Symbols

a	linear acceleration
I_W	moment of inertia of glass body
I_C	moment of inertia of case
I_{WG}	moment of inertia of gyro rotor
K	viscous damping coefficient, Boltzmann constant
C	torsion bar constant
G_V	viscous damping compensation gain
G_T	torsion compensation gain
G_A	acceleration compensation gain
W_{WT}	natural frequency of wheel due to torsion bar suspension
W_{WP}	natural frequency of wheel due to pendulousness
W_α	natural frequency of precession axis due to pendulousness
M_D	disturbing torque on case axis
$M_{D\alpha}$	disturbing torque on gyro precession axis
M_C	control torque
g	acceleration of gravity
g_0	servo position gain
g_1	servo rate gain
g_{-1}	servo integral gain
φ	angle between case and support in reference frame angle on measuring axis
ε	angle of glass body in reference frame
α	gyro precession angle
η_a	accelerometer accuracy
η_T	accuracy of torsion compensation
V	vehicle velocity
	control voltage
V_0	power supply voltage
$\dot{\varphi}$	time derivative
ρ	signal to feedback ratio
M	modulation factor
f_M	frequency of modulation
H	gyro angular momentum
Ω	gyro wheel speed
U	fluid velocity due to temperature gradient
S	distance
T	temperature
t	time
η	Kinematic viscosity
$\bar{a}$	temperature conductivity
r_0	radius of case
r	distance from wall

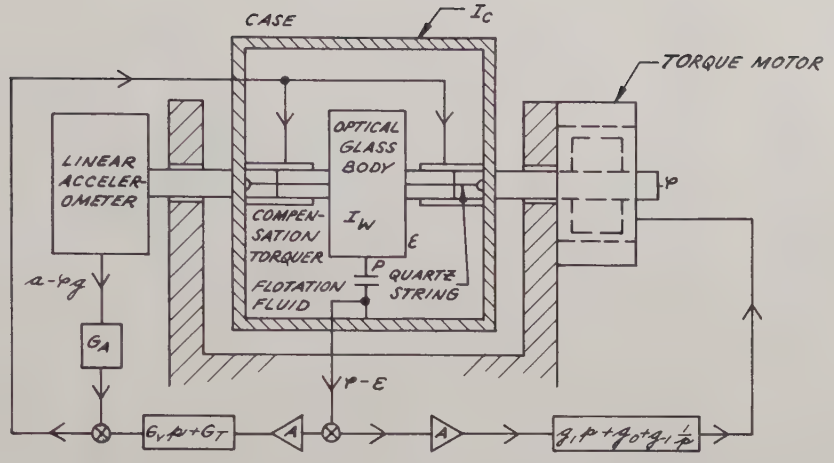


FIG. 1.
SYSTEM SCHEMATIC
OF NONGYROSCOPE
INERTIAL REFERENCE SYSTEM

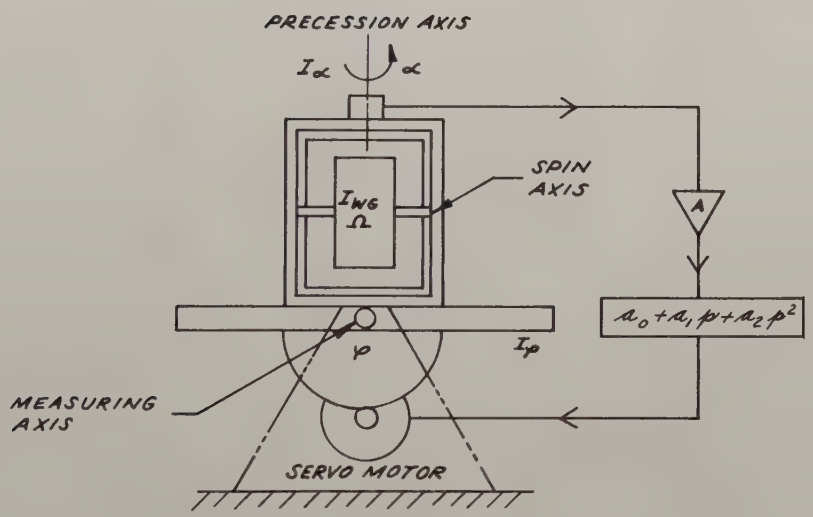


FIG. 2.
SCHEMATIC OF AIR-BEARING
GYRO SYSTEM

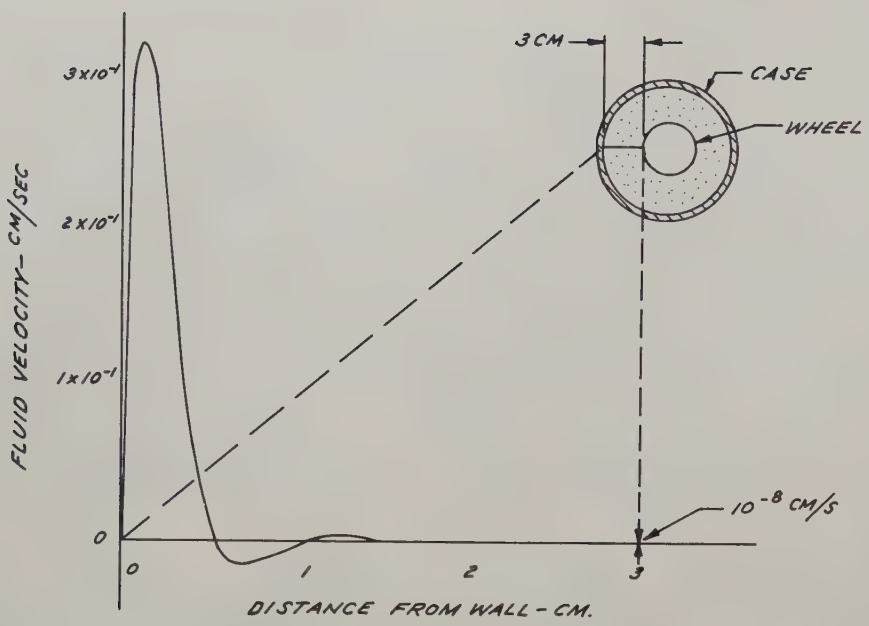
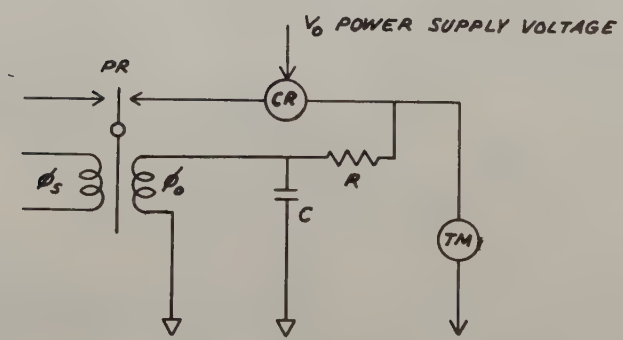
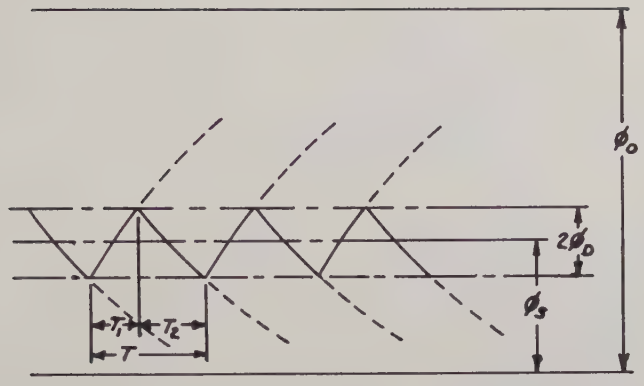


FIG. 3.
VELOCITY PROFILE
DUE TO A TEMPERATURE DISTURBANCE OF: $\Delta T = 0.1°C$



PR POLORIZED RELAY
CR CONTROLLED RECTIFIER
TM TORQUE MOTOR

FIG. 4
SCHEMATIC OF
PULSE MODULATION
CONTROLLER



ϕ_s FLUX DUE TO INPUT SIGNAL
ϕ_o FLUX DUE TO FEEDBACK
ϕ_o FLUX NECESSARY TO OVERCOME DEAD ZONE

FIG. 5
MODULATION PROCESS

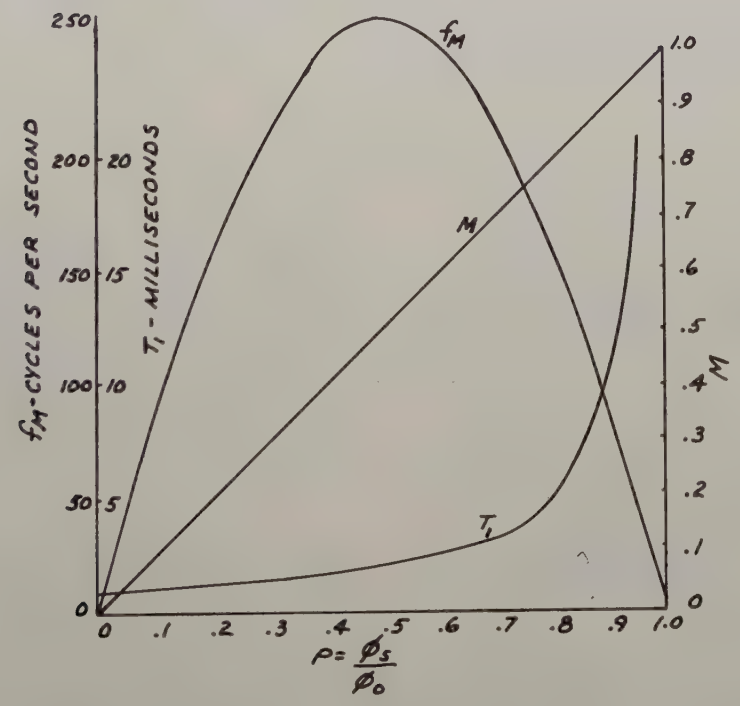


FIG. 6
MODULATION CHARACTERISTICS

SAMPLED DATA DESIGN BY LOG GAIN DIAGRAMS

M. P. Pastel and G. J. Thaler
U. S. Naval Postgraduate School
Monterey, California

Summary

The bilinear transformation $z = (1+w)/(1-w)$ converts a z-transform function G(z) of a sampled-data system into a new function G(w), called the w-transform function, which is a rational function in variable w. This bilinear transformation maps the unit circle on the z-plane onto the imaginary axis of the w-plane. Consequently, it is now possible to readily draw log magnitude and phase diagrams against a frequency scale of the open-loop w-transform function of a sampled-data system by use of asymptotic techniques. Then, by use of a Nichols chart and correlation information available from continuous systems, it is possible to predict the approximate time domain performance. Design by modification of the open-loop transfer function can be made on the diagram in the same manner as employed for continuous systems on the Bode diagram. The resulting w-transform can be converted to its equivalent Laplace transform. The ratio of this transform function and the original Laplace transform function of the system's equipment gives the required compensator. Remote s-plane poles may have to be added to have the compensator physically realizable. Restricting the modifying w-plane poles to lie between (0) and (-1) permits the compensator to be realizable as an RC network.

Introduction

In the last few years methods based on the z-transform have been advanced to aid in the analysis and design of sampled-data feedback control systems.[1,2,3,4] In general, these methods have involved root-locus diagrams or Nyquist diagram techniques. However, when compared to the methods of designing continuous systems these z-transform techniques have two serious limitations.

1. It is not possible to employ the Bode diagram method of analysis in which log magnitude and phase diagrams are plotted against a frequency scale by use of asymptotic techniques.
2. Directly connecting a continuous compensator into the system requires complete recalculation of the system z-transfer function.

It is the purpose of this paper to present an integrated straightforward method of designing sampled-data control systems which will not be subject to the above-mentioned limitations. It will be shown how restrictions on the method will permit utilization of continuous RC networks for the compensator.

Construction of Log Magnitude and Phase Diagrams

The z-transfer function G(z) is obtained from the pulse transfer function G*(s) by making the substitution

$$z = e^{sT} \tag{1}$$

where T is the sampling period. The resulting z-transform is a rational function of z. The conformal mapping according to the relationship expressed by Equation 1 is shown in Figure 1 where ω_s is the sampling frequency. It is significant that the $j\omega$ axis on the s-plane maps into the unit circle on the z-plane. This indicates the difficulty associated with the technique of plotting log magnitude and phase versus a log frequency scale of the z-transfer function as the z variable traverses the unit circle. The various factors of the transform do not sequentially dominate the function and hence the asymptotic plotting techniques are not applicable.

It has been suggested[5] that this difficulty can be circumvented by a change of the independent variable given by the bilinear transformation

$$z = \frac{1 + w}{1 - w} \tag{2}$$

and

$$w = \frac{z - 1}{z + 1} \tag{3}$$

Inasmuch as the transformation is linear, the new function G(w), called the w-transform function, remains a rational function of the variable w. The conformal mapping of the relation in Equation 3 is shown in Figure 2 where it is seen that the unit circle on the z-plane has for its image the imaginary axis on the w-plane. Considering the complex quantity w as containing a real part u and an imaginary part v, a frequency response may be calculated for the transfer function by letting

$$w = jv \tag{4}$$

Since v varies from zero to infinity the asymptotic plotting techniques are directly applicable. Then, from the frequency response, relative stability may be predicted by use of the Nichols chart.

As an illustration of this analysis method consider the sampled-data system shown in Figure 3 with

$$G(s) = \frac{1.34}{s(s + 1)} \tag{5}$$

and having a sampling frequency of 2π radians per second. The corresponding z-transform is

$$G(z) = \frac{0.847z}{(z - 1)(z - .368)} \qquad (6)$$

Substitution of the change of variable of Equation 2 into the above function yields

$$G(w) = \frac{0.671(1 - w^2)}{w(\frac{w}{.462} + 1)} \qquad (7)$$

The log magnitude and phase diagrams for this function against the variable v is shown in Figure 4. Transferring this log magnitude and phase locus to the Nichols chart, as is done in Figure 5, shows a resonance magnification of 43%. The presence of a resonance magnification indicates that a pole of the system function is in the proximity of the imaginary axis. It is evident that the closer the pole to the imaginary axis, the greater the resonance peak and the lower the degree of damping in the transient response. Only limited correlation data is available between frequency response and the output transient response for sampled-data systems.[2,8] However, applying directly the standard correlation [6] for a second order continuous system predicts a peak overshoot of 28% which checks closely with the actual transient peak of 27.6%. The accuracy of the prediction in this case is due to the simplicity of the system and the high sampling frequency. However, less correlation between resonance peak and peak overshoot of the transient response should be expected for systems of more complexity and for systems employing a sampling frequency comparable to the system bandwidth.

Design with the w-Transform

To employ the w-transform and the Bode diagram for analysis and design, it is important to consider some general qualities of w-transfer functions and its linear equivalent, the z-transfer function. The latter function is characterized by the highest power of z in the denominator exceeding that of the numerator by the number of sampling instants having a zero output after the first sampling instant at time equal to zero. Specifically, a Laplace transfer function having a denominator of two degrees or more higher than the numerator will have a corresponding z-transform with the denominator only one degree higher than the numerator. Hence, the factored form of the z-transform will exhibit Z-plane zeros for which there may be no corresponding s-plane image. However, the z-plane poles of the transfer function have direct s-plane images given by Equation 1.

When the substitution given by Equation 2 is employed, the factored G(w) function has the same number of poles as the G(z) function from which it was derived. However, the number of zeros equals the number of poles with all the additional zeros occurring at 1.0. Thus, any Bode diagram of a w-transfer function must satisfy these requirements.

Inspection of Figures 1 and 2 shows that the left hand half of the s-plane maps into the left hand half of the w-plane. Consequently, for physical systems w-plane poles should be chosen in the left hand plane. However, the same restrictions do not apply to the zeros.

Within the restrictions discussed in the preceding paragraphs, an original $G_1(w)$ may be modified on the Bode diagram to satisfy specifications. If, as is usually the case, the compensator is not permitted to have a pole at infinity, the finally selected transfer function $G_2(w)$ must have the same or greater difference between the number of poles and zeros other than those at 1.0 as found in the transfer function $G_1(w)$.

$G_2(w)$ may be converted into its equivalent Laplace transfer function $G_2(s)$ and the transfer function of the compensator $G_c(s)$ found by simple division.

$$G_c(s) = \frac{G_2(s)}{G_1(s)} \qquad (8)$$

Because of the lack of correspondence between the number of s-plane zeros and w-plane zeros, the function $G_c(s)$ derived from Equation 8 will in many cases have more zeros than poles. The requirement for a pole at infinity may be circumvented in such cases by approximating the compensator with remote poles on the negative real s-plane axis, such as is done in the Guillemin direct synthesis procedure for continuous systems.[7] These additional poles have negligible effect on the system response. The final system then has the form shown in Figure 6 with the compensator cascaded directly into the fixed parts of the system.

Compensating with RC network

RC networks are attractive as compensators because of their economy in cost and space as well as ease of assembly. However, to employ only RC networks additional restrictions are required for the final form of the $G_2(w)$ function.

Network synthesis theory lists the following physical realizability conditions for RC transfer functions.

1. The poles are simple and restricted to the negative real axis excluding the origin or infinity.

2. The zeros may be of any order and anywhere.

The first requirement means that added w-poles must be placed on the real w-plane axis from (0.0) to (-1.0). The second requirement gives complete freedom to the selection of w-zeros.

Example

As a complete example of the methods presented in this paper, consider a system having a sampling period of one second and a transfer function

$$G_1(s) = \frac{2.24}{s(s+1)(s+2)} \tag{9}$$

$$G_1(z) = \frac{.448z(z-0.368)}{(z-1)(z-.368)(z-.1354)} \tag{10}$$

The corresponding w-transform is then

$$G_1(w) = \frac{.56(1-w)(w+1)(\frac{w}{2.165}+1)}{w(\frac{w}{.462}+1)(\frac{w}{.761}+1)} \tag{11}$$

Figure 7 shows the Bode diagram for Equation 11 and by use of the Nichols chart in Figure 8 the transient peak overshoot is predicted as 53%. Compensating the system with a phase lag network with a w-plane pole at -0.02 and a w-plane zero at -0.07 and holding the low frequency gain constant yields a new function

$$G_2(w) = \frac{.56(1+w)(1-w)(\frac{w}{.07}+1)(\frac{w}{2.165}+1)}{w(\frac{w}{.02}+1)(\frac{w}{.462}+1)(\frac{w}{.761}+1)} \tag{12}$$

The predicted transient response overshoot is 25%. The actual transient output sequence of samples has a peak overshoot of 27.6%. Consequently, the system as compensated is considered satisfactory. The corresponding Laplace transfer function is then

$$G_2(s) = \frac{.0289(s+.14)(s+22.31)}{s(s+.041)(s+1)(s+2)} \tag{13}$$

Without approximation this leads to a compensation function

$$G_c(s) = \frac{.0129(s+.14)(s+22.31)}{(s+.041)} \tag{14}$$

Equation 14 can be made RC realizable with negligible effect on the system response by adding a pole at (-40). This yields an RC realizable compensation function

$$G_c(s) = \frac{.516(s+.14)(s+22.31)}{(s+.041)(s+40)} \tag{15}$$

This completes the design. A similar procedure can be used with other system functions. Further check on the suitability of the transient response between sampling instants can be made by employing the modified z-transformation integral.[8]

References

1. The Analysis of Sampled-Data Systems, J. R. Ragazzine and L. A. Zadeh. AIEE Transactions Vol. 71, pt. II, 1952.

2. Analysis and Synthesis of Sampled-Data Control Systems, E. I. Jury. AIEE Transactions Vol. 73, pt. I.

3. Root Locus and Transient Response of Sampled-Data Systems, E. I. Jury. AIEE Transactions Vol. 75, pt. II.

4. Root Locus Method of Pulse Transfer Function for Sampled-Data Control Systems, Masahiro Mori, IRE Transactions on Automatic Control, Nov. 1957.

5. Extension of Continuous Data System Design Techniques to Sampled-Data Control Systems, G. W. Johnson, D. P. Lindroff and C. G. A. Nording. AIEE Transactions Vol. 74, Sept. 1955.

6. The Frequency Response Method - A Brief Survey, R. H. Macmillan - Frequency Response (a book), Edited by Oldenburger, The Macmillan Co., 1956.

7. Control System Synthesis (a book), J. G. Truxal, The McGraw Hill Co., 1955.

8. Synthesis and Study of Sampled-Data Control Systems, E. I. Jury, AIEE Transactions pt. II, July 1956.

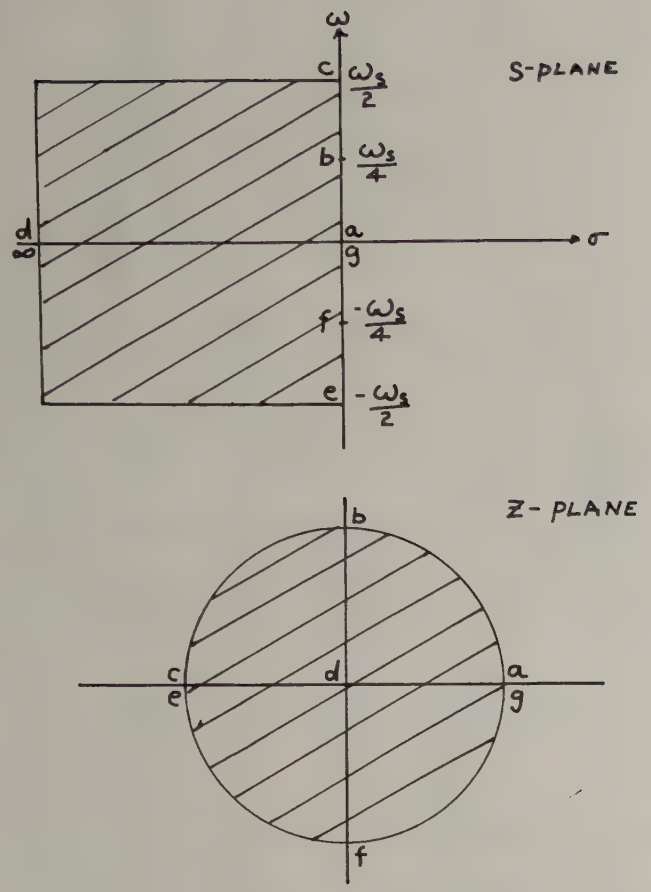


Fig. 1. Conformal mapping of the left-hand half of the s plane into the z plane by the relationship $z = e^{st}$.

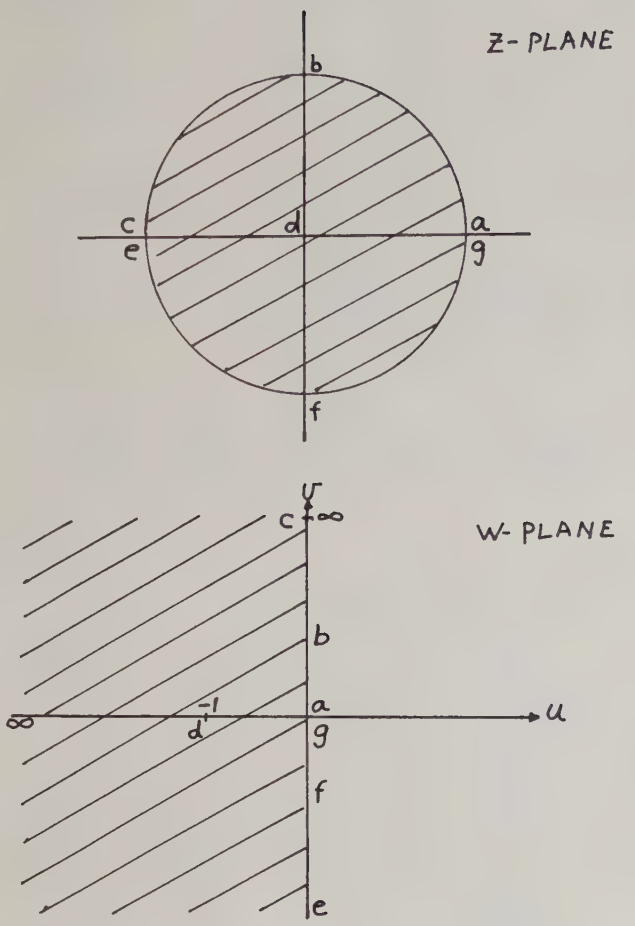


Fig. 2. Conformal mapping of the area with the unit circle on the z plane into the w plane by the relationship $w = \frac{z - 1}{z + 1}$

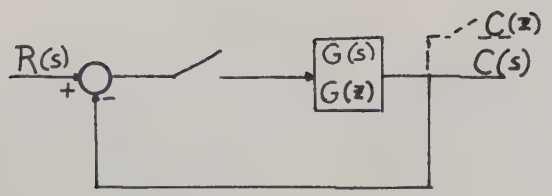


Fig. 3. Sampled-data feedback system.

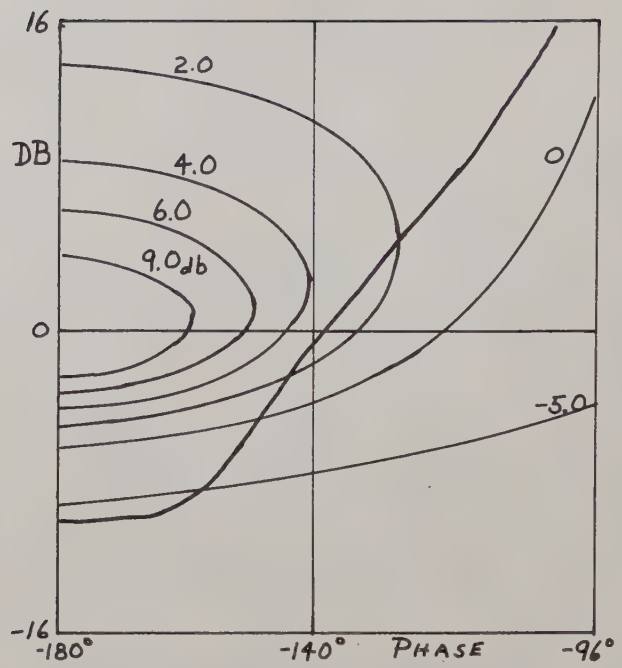


Fig. 5. Nichols Plot for system in example.

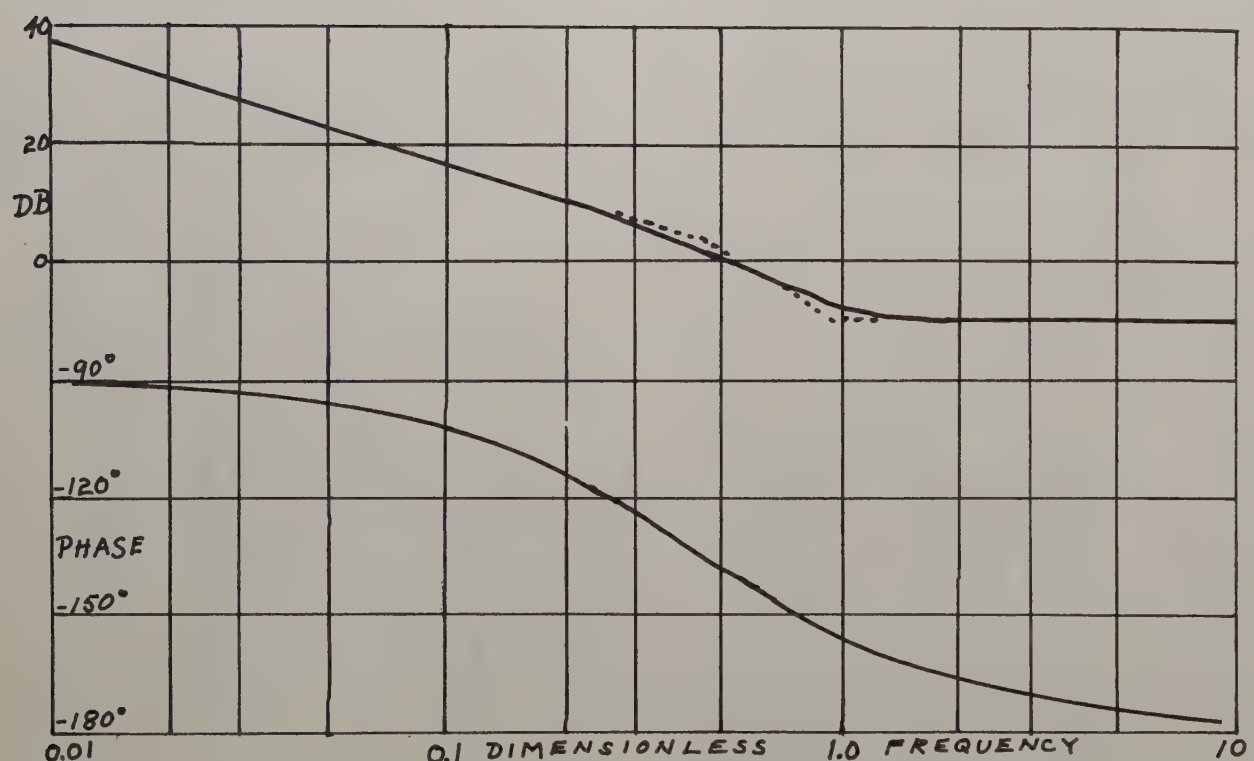


Fig. 4. Attenuation and phase plot for system of example.

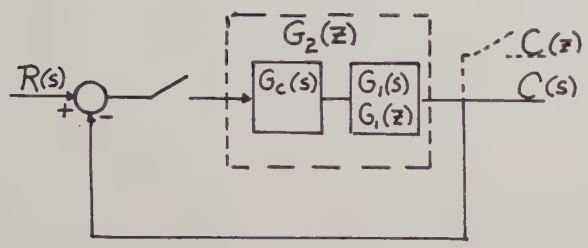


Fig. 6. Compensated sampled-data feedback system.

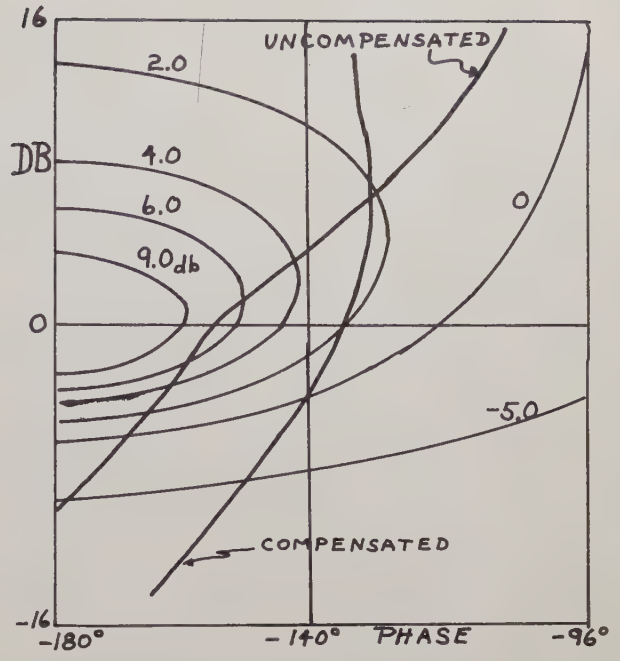


Fig. 8. Nichols Plot for compensated and uncompensated systems.

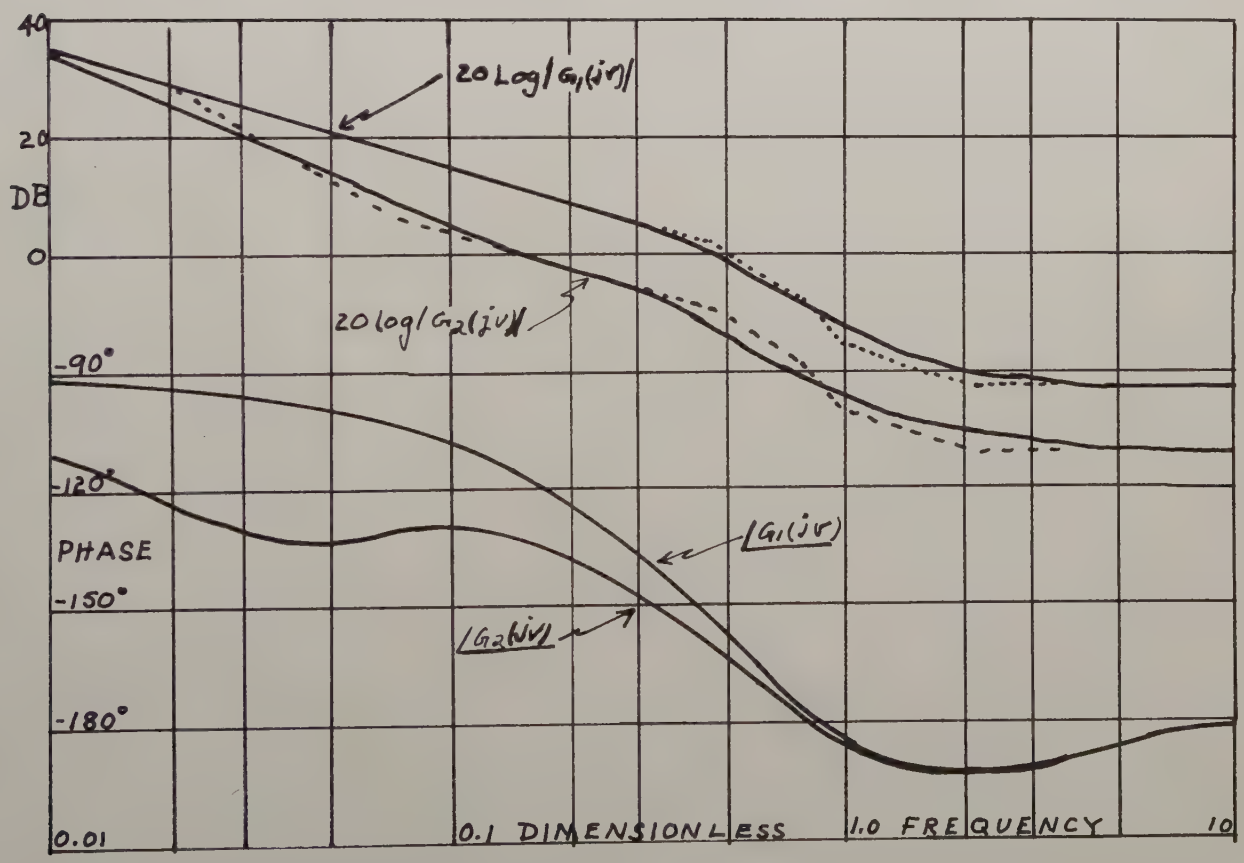


Fig. 7. Attenuation and phase plot for uncompensated and compensated systems.

INSTITUTIONAL LISTINGS

The IRE Professional Group on Automatic Control is grateful for the assistance given by the firms listed below, and invites application for Institutional Listings from other firms interested in the field of Automatic Control.

RAMO-WOOLDRIDGE, DIV. OF THOMPSON RAMO WOOLDRIDGE INC., P.O. Box 90534, Airport Station, Los Angeles 45, Calif.
Electronic Research and Development

PHILCO CORP., GOVERNMENT & INDUSTRIAL DIV., 4700 Wissahickon Ave., Philadelphia 44, Pa.
Transac S-2000 All Transistor, Large-Scale Data-Processing Systems; Transac Computers

The charge for an Institutional Listing is $75.00 per issue or $125.00 for two consecutive issues. Applications for Institutional Listings and checks (made out to the Institute of Radio Engineers, Inc.) should be sent to Mr. L. G. Cumming, Technical Secretary, Institute of Radio Engineers, Inc., 1 East 79th Street, New York 21, N. Y.